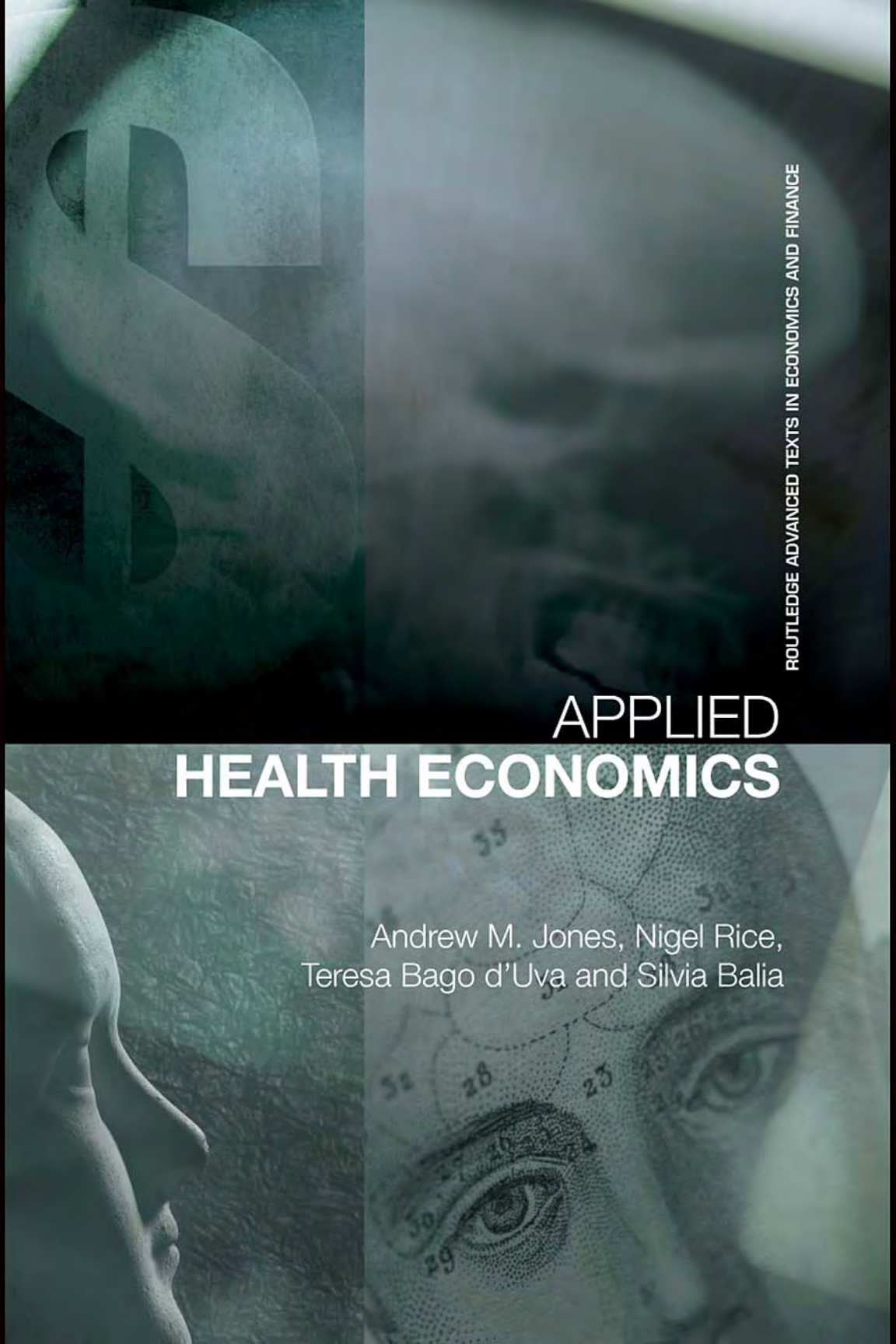

ROUTLEDGE ADVANCED TEXTS IN ECONOMICS AND FINANCE

APPLIED HEALTH ECONOMICS

Andrew M. Jones, Nigel Rice,
Teresa Bago d'Uva and Silvia Balia

Applied Health Economics

Large-scale survey datasets, in particular complex survey designs such as panel data, provide a rich source of information for health economists. They offer the scope to control for individual heterogeneity and to model the dynamics of individual behaviour. However, the measures of outcome used in health economics are often qualitative or categorical. These create special problems for estimating econometric models. The dramatic growth in computing power over recent years has been accompanied by the development of methods that help to solve these problems. This book provides a practical guide to the skills required to put these techniques into practice.

Jones *et al.* illustrate practical applications of these methods using data on health from, among others, the British Health and Lifestyle Survey (HALS), the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP) and the WHO Multi-Country Survey Study (WHO-MCS). Assuming a familiarity with the basic syntax and structure of Stata, this book presents and explains the statistical output using empirical case studies rather than general theory.

A distinctive feature of the text is the way that it brings together theory and practice. This book will be of great benefit to applied economists, as well as advanced undergraduate and post-graduate students in health economics and applied econometrics.

Andrew M.Jones is Professor of Economics and Director of the Graduate Programme in Health Economics at the University of York.

Nigel Rice is Reader in Health Economics at the University of York.

Teresa Bago d'Uva is an Assistant Professor at the Department of Economics, Erasmus University.

Silvia Balia is an Assistant Professor at the Department of Economic and Social Research, University of Cagliari.

Routledge Advanced Texts in Economics and Finance

Financial Econometrics
Peijie Wang

Macroeconomics for Developing Countries 2nd edition
Raghbendra Jha

Advanced Mathematical Economics
Rakesh Vohra

Advanced Econometric Theory
John S.Chipman

Understanding Macroeconomic Theory
John M.Barron, Bradley T.Ewing and Gerald J.Lynch

Regional Economics
Roberta Capello

Mathematical Finance
Core theory, problems and statistical algorithms
Nikolai Dokuchaev

Applied Health Economics
Andrew M.Jones, Nigel Rice, Teresa Bago d'Uva and Silvia Balia

Applied Health Economics

Andrew M.Jones, Nigel Rice, Teresa Bago d'Uva and Silvia Balia

Routledge
Taylor & Francis Group

LONDON AND NEW YORK

First published 2007
by Routledge
2 Park Square, Milton Park, Abingdon OX 14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group,
an informa business

This edition published in the Taylor & Francis e-Library, 2007.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.ebookstore.tandf.co.uk."

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Applied health economics/Andrew M.Jones—[et al.]
p. cm.—(Routledge advanced texts in economics and finance; 8)
Includes bibliographical references and index.
1. Medical economics. I. Jones, Andrew M., 1960–.
RA410.5.A66 2007
338.4′73621—dc22
2006028492

ISBN 0-203-97230-9 Master e-book ISBN

ISBN10:0-415-39771-5 (hbk)
ISBN10:0-415-39772-3 (pbk)
ISBN10:0-203-97230-9 (ebk)
ISBN13:978-0-415-39771-1 (hbk)
ISBN13:978-0-415-39772-8 (pbk)
ISBN13:978-0-203-97230-4 (ebk)

Contents

Illustrations

Figure

Table

Preface

Large-scale survey datasets, in particular complex survey designs such as panel data, provide a rich source of information for health economists. They offer the scope to control for individual heterogeneity and to model the dynamics of individual behaviour. However, the measures of outcome used in health economics are often qualitative or categorical. These create special problems for estimating econometric models. The dramatic growth in computing power over recent years has been accompanied by the development of methods that help to solve these problems. The purpose of this book is to provide a practical guide to the skills required to put these techniques into practice.

Practical applications of the methods are illustrated using data on health from, among others, the British Health and Lifestyle Survey (HALS), the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP) and the WHO Multi-Country Survey Study (WHO-MCS). There is a strong emphasis on applied work, illustrating the use of relevant computer software with code provided for Stata (http://www.stata.com/). Familiarity with the basic syntax and structure of Stata is assumed. The Stata code and extracts from the statistical output are embedded directly in the main text and explained as we go along. The command lines appear in the same format that they are recorded in the Stata log file, prefixed by '•', for example:

The Stata output appears alongside in a smaller font. The code presented in this book can be downloaded from the web at the homepage of the Health-Econometrics and Data Group, http://www.york.ac.uk/res/herc/hedg.html.

We do not attempt to provide a review of the extensive health economics literature that makes use of econometric methods (for a survey of the pre-2000 literature see Jones (2000) and for a collection of papers see Jones and O'Donnell (2002)). Instead, the book is built around empirical case studies, rather than general theory, and the emphasis is on learning by example. We present a detailed dissection of methods and results of some recent research papers written by the authors and our colleagues. Relevant methods are presented alongside the Stata code that can be used to implement them, and the empirical results are discussed as we go along. To our knowledge, no comparable text exists. There are health economics texts and there are econometrics texts but these tend to focus on theory rather than application and tend not to bring the two disciplines together for the benefit of applied economists. The emphasis is on hands-on empirical analysis: the kind of thing that econometric texts tend to neglect. The closest in spirit is Angus Deaton's (1997) excellent book on the analysis of household surveys, but that emphasizes issues in the economics of development, poverty and welfare rather than health. A general knowledge of microeconometric methods is assumed. For more details readers can refer to texts such as Baltagi (2005), Cameron and Trivedi (2005), Greene (2003) and Wooldridge (2002b).

• use "c:\stata\data\bhps.dta", clear

As the book is built around case studies, and these reflect the particular interests of the authors, we do not claim to cover the full diversity of topics within applied health economics. However, we hope that these examples will provide guidance and inspiration for those working on other topics within the field who want to make use of econometric methods. The book is primarily aimed at advanced undergraduates and postgraduates in health economics, along with health economics researchers in academic, government and private sector organizations who want to learn more about empirical research methods. In addition, the book may be used by other applied economists, in areas such as labour and environmental economics, and by health and social statisticians.

Acknowledgements

Data from the British Household Panel Survey (BHPS) were supplied by the UK Data Archive. Neither the original collectors of the data nor the archive bear any responsibility for the analysis or interpretations presented here. The European Community Household Panel Users' Database (ECHP), version of December 2003, was supplied by Eurostat. Data from the Health and Lifestyle Survey (HALS) were supplied by the UK Data Archive. Neither the original collectors of the data nor the archive bear any responsibility for the analysis or interpretation presented here. We are grateful to Statistics Canada for access to the National Population Health Survey (NPHS) data. We thank the World Health Organization for providing access to the WHO Multi-Country Survey Study (WHO-MCS) data.

We are very grateful to all of the co-authors of the joint work that we use as case studies: Somnath Chatterji, Paul Contoyannis, Martin Forster, Cristina Hernández-Quevedo, Xander Koolman, Maarten Lindeboom, Owen O'Donnell, Jennifer Roberts and Eddy van Doorslaer. The specific papers that are adapted for the case studies are:

Bago d'Uva, T. (2006) 'Latent class models for health care utilisation', *Health Economics,* 15:329–343.

Bago d'Uva, T., van Doorslaer, E., Lindeboom, M., O'Donnell, O. and Chatterji, S. (2006) 'Does reporting heterogeneity bias the measurement of health disparities?', HEDG Working Paper 06/03, University of York.

Balia, S. and Jones, A.M. (2005) 'Mortality, Lifestyle and Socio-Economic Status', HEDG Working Paper 05/02, University of York.

Contoyannis, P., Jones, A.M. and Rice, N. (2004) 'Simulation-based inference in dynamic panel probit models: an application to health', *Empirical Economics,* 29:49–77.

Contoyannis, P., Jones, A.M. and Rice, N. (2004) 'The dynamics of health in the British Household Panel Survey', *Journal of Applied Econometrics,* 19:473–503.

Contoyannis, P. and Rice, N. (2001) 'The impact of health on wages: Evidence from the British Household Panel Survey', *Empirical Economics,* 26:599–622.

Forster, M. and Jones, A.M. (2001) 'The role of tobacco taxes in starting and quitting smoking: duration analysis of British data', *Journal of the Royal Statistical Society Series A,* 164:517–547.

Jones, A.M., Koolman, X. and Rice, N. (2006) 'Health-related non-response in the BHPS and ECHP: using inverse probability weighted estimators in nonlinear models', *Journal of the Royal Statistical Society Series A,* 169:543–569.

Rice, N., Roberts, J. and Jones, A.M. (2006) 'Sick of work or too sick to work? Evidence on health shocks and early retirement form BHPS', HEDG Working Paper 06/13, University of York.

van Doorslaer, E. and Jones, A.M. (2003) 'Inequalities in self-reported health: validation of a new approach to measurement', *Journal of Health Economics,* 22:61–87.

A draft of the book was used as teaching material for a short course entitled ‘Applied Health Economics’ that was hosted by the Health, Econometrics and Data Group (HEDG) at the University of York, 19–30 June 2006. This course was part of the Marie Curie Training Programme in Applied Health Economics. We are grateful for the input from other members of HEDG who were involved with the course: Cristina Hernández Quevedo, Eugenio Zuccheli, Silvana Robone, Pedro Rosa Dias and Rodrigo Moreno Serra. We should also like to thank the course participants for their valuable feedback on the material.

Finally, we should like to thank Rob Langham at Routledge for encouraging us to take on this project and for his patience and support in seeing it through.

Introduction

This book provides a practical guide to applied health economics. It is built around a series of case studies that are based on recent research. The first, which runs through the book, explores the dynamics of self-reported health in the British Household Panel Survey (BHPS). The aim is to investigate socioeconomic gradients in health, persistence of health problems and the difficulties created by sample attrition in panel data (Contoyannnis, Jones and Rice 2004; Jones, Koolman and Rice 2006). The data for this and all the other case studies are introduced in Chapter 1, which also introduces some general principles of survey design. Chapter 2 uses the BHPS sample to show how descriptive techniques, including graphs and tables, can be used to summarize and explore the raw data and provide an intuitive understanding of how variables are distributed and associated with each other.

Distributional analysis is taken a step further in Chapter 3, which also introduces some basic regression models for cross-section surveys: linear, ordered and interval regressions. The chapter uses Canadian data, from the National Population Health Survey (NPHS), on self-reported health and an index of health-related quality of life, the 'HUI' (van Doorslaer and Jones 2003). These kinds of subjective and self-reported measures of health raise questions of reliability. Chapter 4 explores the issue of reporting bias using Indian data from the World Health Organization's Multi-Country Survey Study (WHO-MCS). The standard ordered probit model is extended to include applications of the generalized ordered model and the 'HOPIT'. These exploit hypothetical Vignettes' to deal with reporting bias (Bago d'Uva *et al.* 2006).

Lifestyle factors, such as smoking, drinking and exercise, are thought to have an influence on health. But these health-related behaviours are individual choices that are themselves influenced by, often unobservable, individual characteristics such as time preference rates. Chapter 5 uses data from the Health and Lifestyle Survey (HALS) to show how the multivariate probit model can be used to model mortality, morbidity and lifestyles jointly, taking account of the problem of unobservables (Balia and Jones 2005). This illustrates the kind of models that can be applied to categorical data in cross-section surveys.

Part III moves from cross-sectional data to longitudinal data, in particular to duration analysis. There are two types of duration data: continuous and discrete time. Chapter 6 takes the analysis of HALS a step further by estimating continuous time duration models for initiation and cessation of smoking and for the risk of death (this draws on earlier work by Forster and Jones 2001). Chapter 7 illustrates convenient methods for discrete-time duration analysis. The BHPS is used to investigate the extent that 'health shocks' constitute a factor that leads to early retirement.

Longitudinal data is the focus of Part IV, which presents linear and non-linear panel data regression methods. Linear models are covered in Chapter 8, where BHPS data are used to estimate classical Mincerian wage equations that are augmented by measures of self-reported health (Contoyannis and Rice 2001). Chapter 9 stays with the BHPS but

moves to non-linear dynamic specifications (Contoyannis, Jones and Rice 2004). The outcome of interest is a binary measure of health problems and the focus is on socioeconomic gradients in health. Chapter 10 continues this analysis but shifts the emphasis to the potential problems created by sample attrition in panel data (Jones, Koolman and Rice 2006). The chapter shows how to test for attrition bias and illustrates how inverse probability weights provide one way of dealing with the problem.

Finally, Chapter 11 turns to health-care utilization, exploring data on specialist visits from the European Community Household Panel (ECHP). Health-care utilization is most frequently modelled using count data regressions. The chapter reviews and applies standard methods and also introduces recent developments of the literature that use a latent class specification (Bago d'Uva 2006).

Part I
Data description

1
Data and survey design

This chapter introduces each of the datasets that are used in the practical case studies throughout the book. It discusses some important features of survey design and focuses on the variables that are of particular interest to health economists.

1.1 THE HEALTH AND LIFESTYLE SURVEY (HALS)

The sample

The Health and Lifestyle Survey (HALS) is an example of a health interview survey. Aspects of the survey are used in Chapters 5 and 6. The HALS was designed as a representative survey of adults in Great Britain (see Cox *et al.* 1987 and 1993). The population surveyed comprised individuals aged 18 and over living in private households. In principle, each individual should have an equal probability of being selected for the survey. This allows the data to be used to make inferences about the underlying population. HALS was designed originally as a cross-section survey with one measurement for each individual. It was carried out between the autumn of 1984 and the summer of 1985, and information was collected in three stages:

- A one-hour face-to-face interview, which collected information on experience and attitudes towards health and lifestyle along with general socioeconomic information.
- A nurse visit to collect physiological measures and indicators of cognitive function, such as memory and reasoning.
- A self-completion postal questionnaire to measure psychiatric health and personality.

The HALS is an example of a clustered random sample. The intention was to build a representative random sample of this population but without the excessive costs of collecting a true random sample. Addresses were randomly selected from electoral registers using a three-stage design. First 198 electoral constituencies were selected with the probability of selection proportional to the population of each constituency. Then two electoral wards were selected for each constituency and, finally, 30 addresses per ward. Then individuals were randomly selected from households. This selection procedure gave a target of 12,672 interviews.

Some of the addresses from the electoral register proved to be inappropriate as they were in use as holiday homes or business premises, or were derelict. This number was relatively small and only 418 addresses were excluded, leaving a total of 12,254

individuals to be interviewed. The response rate fell more dramatically when it came to success in completing these interviews; 9,003 interviews were completed. This is a response rate of 73.5%. In other words, there was a 1 in 4 chance that an interview was not completed.

The overall response rate is fairly typical of general population surveys. Understandably, the response rate declines for the subsequent nurse visit and postal questionnaire. The overall response rate for those individuals who completed all three stages of the survey is only 53.7%. To get a sense of how well the sample represents the population it can be compared to external data sources. The most comprehensive of these is the population census, which is collected every ten years. Comparison with the 1981 census suggests that the final sample under-represents those with lower incomes and lower levels of education. In general, it is important to bear this kind of unit non-response in mind when analysing any survey data.

The longitudinal follow-up

The HALS was originally intended to be a one-off cross-sectional survey. However, HALS also provides an example of a longitudinal, or panel, dataset. In 1991/92, seven years on from the original survey, the HALS was repeated. This provides an example of repeated measurements, where the same individuals are re-interviewed. Panel data provide a powerful enhancement of cross-sectional surveys that allows a deeper analysis of heterogeneity across individuals and of changes in individual behaviour over time. However, because of the need to revisit and interview individuals repeatedly the problems of unit non-response tend to be amplified. Of the original 9,003 individuals who were interviewed at the time of the first HALS survey 808 (9%) had died by the time of the second survey, 1,347 (15%) could not be traced and 222 were traced but could not be interviewed, either because they had moved overseas or they had moved to geographic areas that were out of the scope of the survey. These cases are examples of attrition—individuals who drop out of a longitudinal survey.

The deaths data

HALS provides an example of a cross-sectional survey (HALS1) and panel data (HALS1 & 2). Also it provides a longitudinal follow-up of subsequent mortality and cancer cases among the original respondents. These deaths data can be used for survival analysis. Most of the 9003 individuals interviewed in HALS1 have been *flagged* on the NHS Central Register. In June 2005 the fifth death revision and the second cancer revision were completed. The flagging process was quite lengthy because it required several checks in order to be sure that the flagging registrations were related to the person previously interviewed. About 98% of the sample has been flagged. Deaths account for some 27% of the original sample. This information is used in Chapter 6 for a duration analysis of mortality rates.

1.2 THE BRITISH HOUSEHOLD PANEL SURVEY (BHPS)

The sample

The British Household Panel Survey (BHPS) is a longitudinal survey of private households in Great Britain that provides rich information on socio-demographic and health variables. While HALS has only two waves of panel data, the BHPS has repeated annual measurements from 1991 to the present and is an ongoing survey. This provides more scope for longitudinal analysis. The BHPS is used in Chapters 2, 7, 8, 9 and 10.

The BHPS was designed as an annual survey of each adult (aged 16+) member of a nationally representative sample of more that 5,000 households, with a total of approximately 10,000 individual interviews. The first wave of the survey was conducted between 1 September 1990 and 30 April 1991. The initial selection of households for inclusion in the survey was performed using a two-stage clustered systematic sampling procedure designed to give each address an approximately equal probability of selection (Taylor *et al.* 1998). The same individuals are re-interviewed in successive waves and, if they split off from their original households, are also re-interviewed along with all adult members of their new households.

Measures of health

One measure of health outcomes that is available in the BHPS, and many other general surveys, is self-assessed health (SAH), defined by a response to: 'Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?' SAH should therefore be interpreted as indicating a perceived health status relative to the individual's concept of the 'norm' for their age group. SAH has been used widely in previous studies of the relationship between health and socioeconomic status (e.g., Ettner 1996; Deaton and Paxson 1998; Smith 1999; Benzeval *et al.* 2000; Salas 2002; Adams *et al.* 2003; Frijters *et al.* 2003; Contoyannis, Jones and Rice 2004) and of the relationship between health and lifestyles (e.g., Kenkel 1995; Contoyannis and Jones 2004). SAH is a simple subjective measure of health that provides an ordinal ranking of perceived health status. However it has been shown to be a powerful predictor of subsequent mortality (see e.g., Idler and Kasl 1995; Idler and Benyamini 1997) and its predictive power does not appear to vary across socioeconomic groups (see e.g., Burström and Fredlund 2001). Socioeconomic inequalities in SAH have been a focus of research (see e.g., van Doorslaer *et al.* 1997; van Ourti 2003; van Doorslaer and Koolman 2004) and have been shown to predict inequalities in mortality (see e.g., van Doorslaer and Gerdtham 2003). Categorical measures of SAH have been shown to be good predictors of subsequent use of medical care (see e.g., van Doorslaer *et al.* 2000; van Doorslaer *et al.* 2004).

Unfortunately there was a change in the wording of the SAH question at wave 9 of the BHPS. For waves 1–8 and 10 onwards, the SAH variable represents 'health status over

the last 12 months'. However, the SF-36 questionnaire was included in wave 9. In this questionnaire, the SAH variable for wave 9 represents 'general state of health', using the question: 'In general, would you say your health is: excellent, very good, good, fair, poor?' Note that the question is not framed in terms of a comparison with people of one's own age and the response categories differ from the other waves. Item non-response is greater for SAH at wave 9 than for the other waves and these factors would complicate the analysis of non-response rates. Hernández-Quevedo *et al.* (2004) have explored the sensitivity of models of SAH to this change in the wording.

Other indicators of morbidity are available in the BHPS. The variable HLLT measures self-reported functional limitations and is based on the question 'does your health in any way limit your daily activities compared to most people of your age?' Respondents are left to define their own concepts of health and their daily activities. In contrast, for the variable measuring specified health problems (HLPRB), respondents are presented with a prompt card and asked, 'do you have any of the health problems or disabilities listed on this card?' The list is made up of problems with arms, legs, hands, etc; sight; hearing; skin conditions/allergies; chest/breathing; heart/ blood pressure; stomach/digestion; diabetes; anxiety/depression; alcohol/drug related; epilepsy; migraine and other (cancer and stroke were added as separate categories in wave 11). Also respondents are asked to report whether they are registered as a disabled person (HLDSBL).

Socioeconomic status

The analysis of the BHPS data discussed in subsequent chapters often focuses on socioeconomic gradients in health. Two main dimensions of socioeconomic status are included in our analyses: income and education. Income is measured as equivalized and RPI-deflated annual household income (INCOME). In our analysis this variable is often transformed to natural logarithms to allow for concavity of the relationship between health and income (e.g., Ettner 1996; Frijters *et al.* 2003; van Doorslaer and Koolman 2004; Contoyannis, Jones and Rice 2004). Education is measured by the highest educational qualification attained by the end of the sample period in descending order of attainment (DEGREE, HND/A, O/CSE). NO-QUAL (no academic qualifications) is the reference category for the educational variable. In addition to income and education, variables are included to reflect individuals' demographic characteristics and stage of life: age, ethnic group, marital status and family composition. Marital status distinguishes between WIDOW, SINGLE (never married) and DIVORCED/ SEPARATED, with married or living as a couple as the reference category. Similarly, we include an indicator of ethnic origin (NON-WHITE), the number of individuals living in the household including the respondent (HHSIZE), and the numbers of children living in the household at different ages (NCH04, NCH511, NCH1218). Age is included as a fourth-order polynomial, (AGE, AGE2=AGE^2/100, AGES=AGE^3/10000, AGE4= AGE^4/1000000), where the higher-order terms are rescaled to avoid computational problems in the estimation routines.

1.3 THE EUROPEAN COMMUNITY HOUSEHOLD PANEL (ECHP)

The sample

The European Community Household Panel User Database (ECHP-UDB) adds an international dimension and allows a comparison across countries as well as across time. It is used in Chapter 11.

The ECHP was designed and coordinated by Eurostat, the European Statistical Office, and is a standardized multi-purpose annual longitudinal survey carried out at the level of the pre-enlargement European Union (EC-15). More information about the survey can be found in Peracchi (2002). The survey is based on a standardized questionnaire that involves annual interviewing of a representative panel of households and individuals of 16 years and older in each of the participating EU member states. It covers a wide range of topics including demographics, income, social transfers, health, housing, education and employment. Data are used for the following 14 member states of the EU for the full number of waves available for each: Austria (waves 2–8), Belgium (1–8), Denmark (1–8), Finland (3–8), France (1–8), Germany (1–3), Greece (1–8), Ireland (1–8), Italy (1–8), Luxembourg (1–3), Netherlands (1–8), Portugal (1–8), Spain (1–8) and the United Kingdom (1–3). Sweden did not take part in the ECHP although the living conditions panel is included with the UDB. The ECHP-UDB also includes comparable versions of the BHPS and German Socioeconomic Panel (GSOEP).

Measures of health

In the ECHP self-assessed general health status (SAH) is measured as either very good, good, fair, poor or very poor. Unlike the BHPS, respondents are not asked to compare themselves with others of the same age. In France a six-category scale was used but this is recoded to the five-category scale in the ECHP-UDB. Responses are also available for the question 'Do you have any chronic physical or mental health problem, illness or disability? (yes/no)' and if so 'Are you hampered in your daily activities by this physical or mental health problem, illness or disability? (no; yes, to some extent; yes, severely)'.

Socioeconomic status

The ECHP includes a comprehensive set of information on household and personal income, broken down by source. In our analysis the principal income measure is disposable household income per equivalent adult, using the modified OECD equivalence scale (giving a weight of 1.0 to the first adult, 0.5 to the second and each subsequent person aged 14 and over, and 0.3 to each child aged under 14 in the household). Total household income includes all net monetary income received by the household members during the reference year. Education is measured by the highest level of general or higher education completed, i.e. third level education (ISCED 5–7), second stage of secondary level education (ISCED 3) or less than second stage of secondary education (ISCED

0–2). Marital status distinguishes between married/living in consensual union, separated/divorced, widowed and unmarried. Activity status includes employed, self-employed, student, unemployed, retired, doing housework and 'other economically inactive'. Region of residence uses the EU's NUTS 1 level (Nomenclature of Statistical Territorial Units) except for countries where such information was withheld for confidentiality reasons (Netherlands, Germany) or because the country is too small (Denmark, Luxembourg).

1.4 THE CANADIAN NATIONAL POPULATION HEALTH SURVEY (NPHS)

The sample

The data used in Chapter 3 are taken from the first wave (in 1994–1995) of the Canadian National Population Health Survey (NPHS). The target population of the NPHS includes household residents in all provinces, with the exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas of Ontario and Quebec. A total of 26,430 households were selected for the survey. In each household, a randomly selected household member, aged 12 years or older, was selected for a more in-depth interview. This interview included questions on health status, risk factors, and demographic and socioeconomic information.

Health variables

The two key variables for our purposes are self-assessed health (SAH) and health status as measured by the Health Utility Index (HUI). As part of the in-depth component of the NPHS, respondents were asked: 'In general, how would you say your health is?' The response categories were excellent, very good, good, fair and poor. Also, each respondent was assigned a Health Utility Index score based on their response to the questions of the eight-attribute Health Utility Index Mark III health status classification system. The Health Utility Index is a generic health status index, developed at McMaster University, that measures both quantitative and qualitative aspects of health (Torrance *et al.* 1995 and 1996; Feeny *et al.* 1995). It provides a description of an individual's overall functional health, based on eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain. The Health Utility Index assigns a single numerical value, between zero and one, for all possible combinations of levels of these eight self-reported health attributes. A score of one indicates perfect health. The Health Utility Index also embodies the views of society concerning health status, inasmuch as preferences about various health states are elicited from a representative sample of individuals.

Socioeconomic variables

Total income before taxes and deductions, as measured in the NPHS, is a categorical variable with 11 response categories. For the purposes of our application, the two lowest income groups—no income and less than Can$5,000—were combined into one group,

thus reducing the number of income categories from 11 to 10. The midpoint of each income category was then attributed to all households in that category and subsequently divided by an equivalence factor equal to (number of household members)$^{0.5}$, to adjust for differences in household size. The income values assigned for the top and bottom groups were $2,500 and $87,500 respectively. Other health determinants included in the analysis are the following: (i) Education level; the highest level of general or higher education completed is available at three levels: recognized third level education (ISCED 5–7), second stage of secondary level of education (ISCED 3) and less than second stage of secondary education (ISCED 0–2); (ii) Marital status distinguishes between married, separated/divorced, widowed and unmarried (including co-habiting); (iii) Activity status includes employed, self-employed, student, unemployed, retired, housework and 'other economically inactive'.

1.5 THE WHO MULTI-COUNTRY SURVEY STUDY (WHO-MCS)

The data used in Chapter 4 are from the WHO Multi-Country Survey Study on Health and Responsiveness 2000–2001 (WHO-MCS), which covered 71 adult populations in 61 countries. Üstün *et al.* (2003) provide a comprehensive report on the goals, design, instrument development and execution of this survey. Individuals were asked to report their health in each of six health domains (mobility, cognitive functioning, affective behaviour, pain or discomfort, self-care and usual activities). In addition, a sub-sample of individuals were asked to rate a set of anchoring vignettes describing fixed ability levels in each health domain. The general idea is to use the responses to these vignettes to identify reporting heterogeneity. Assessments of the individuals' own health, by domain, can then be calibrated against the vignettes, with the aim of purging reporting heterogeneity and giving interpersonally comparable health measures. In Chapter 4 we model the WHO-MCS data on affective behaviour for an Indian state (Andhra Pradesh).

1.6 OVERVIEW

All of the datasets used in this book are examples of surveys that are designed to be representative of a specified population. Normally these are collected using multi-stage clustered random samples, for convenience and economy. The simplest design is a cross-sectional survey in which each individual is measured just once. This may involve face-to-face interviews, medical examinations, telephone interviews or postal questionnaires. Repeated measurements of the same individuals give longitudinal, or panel, data. This provides more scope for analysis of individual heterogeneity and dynamic models.

2
Describing the dynamics of health

2.1 INTRODUCTION

Contoyannis, Jones and Rice (2004) use eight waves of the British Household Panel Survey (BHPS) to model the dynamics of self-assessed health (SAH): this paper forms the basis for the case study reported in this chapter and in Chapters 9 and 10. The main focus of their paper is on the observed persistence in reported health and an assessment of whether this is due to state dependence or to unobservable individual heterogeneity. The paper also provides evidence on the socioeconomic gradient in health and explores whether health-related attrition is an issue for this kind of analysis. The econometric analysis of the BHPS is discussed in more detail in Chapters 9 and 10 below. This chapter concentrates on some preliminary descriptive analysis of the BHPS data and explains the Stata code that can be used to do graphical analysis and to prepare tables of summary statistics.

In this analysis we use both *balanced samples* of respondents, for whom information on all the required variables is reported at each of the eight waves used here, and *unbalanced samples,* which exploit all available observations for wave 1 respondents. Neither sample includes new entrants to the BHPS; the samples only track all of those who were observed at wave 1. In this sense, the analysis treats the sample as a cohort consisting of all those present at wave 1. To be included in the analysis individuals must be 'original sample members' (OSMs) who were aged 16 or over and who provided a valid response for the health measures at wave 1. Our broad definition of non-response encompasses all individuals who are missing at subsequent waves.

The first step is to load the Stata data file, called bhps. dta, that contains the relevant BHPS variables:

• use "c:\stata\data\bhps.dta", clear

Then a log file, bhps .log, is opened to store a permanent record of the results:

• capture log close
• log using "c:\stata\data\bhps.log", replace

As this is a panel dataset it is useful to specify new variables that contain the individual (i) and time (t) indices. These can be used to sort the data prior to analysis:

• iis pid
• tis wavenum
• sort pid wavenum

The BHPS includes missing data owing to both unit and item non-response, so not all individuals in our dataset are observed at every wave. As described above, this gives two options for the analysis: using the unbalanced panel, that includes all available observations, or the balanced panel, that restricts the sample to those individuals who have a complete set of data for all of the waves. The following commands provide a simple way of creating indicator variables for whether or not individuals are in the balanced panel and in the unbalanced panel. These indicators can be used to select the sample in the subsequent estimation commands and also play a role in the analysis of attrition, as discussed in Chapter 10. The commands work by first running a model that includes all the variables that are relevant for subsequent estimation in the list of dependent and independent variables. Here we use a pooled ordered probit (oprobit) but the particular form of the model is not important. The model is run quietly as we are not interested in the regression output *per se:*

• quietly oprobit hlstat widowed nvrmar divsep deghdeg hndalev ocse hhsize nch04 nch511 nch1218 age age2 age3 age4 yr9293 yr9394 yr9495 yr9596 yr9697 yr9798 lninc mlninc mwid mnvrmar mdivsep mhhsize mnch04 mnch511 mnch1218, cluster(pid)

Having run the model we can exploit the saved result e (sample), which holds an indicator of whether an observation has been used in the preceding estimation command. We use this to create an indicator of whether an observation is in the estimation sample or not:

• gen insampm=0
  • recode insampm 0=1 if e (sample)

Then the data are sorted by individual and wave identifiers and a new variable (Ti) is generated that counts the number of waves for which each individual is observed:

• sort pid wavenum
  • gen constant=1
  • by pid: egen Ti=sum (constant) if insampm==1

Using this new variable it is possible to create indicators of whether an individual appears in the next wave (nextwavem) and for whether they appear in the balanced panel (allwavesm). These variables are used in simple tests of attrition that are described in Chapter 10:

• sort pid wavenum
  • by pid: gen nextwavem=insampm [_n+1]
  • gen allwavesm=.
  • recode allwavesm.=0 if Ti~=8
  • recode allwavesm.=1 if Ti==8
  • gen numwavesm=.
  • replace numwavesm=Ti

2.2 GRAPHICAL ANALYSIS

Now we move on to show the Stata code that produces the graphical analysis of self-assessed health (SAH) from Contoyannis, Jones and Rice (2004). First, it is useful to attach some meaningful labels to describe the categorical responses to the question:

• label variable sahex “hlstat=excellent”
• label variable sahgood “hlstat=good”
• label variable sahfair “hlstat=fair”
• label variable sahpoor “hlstat=poor”
• label variable sahvpoor “hlstat=very poor”

Contoyannis, Jones and Rice (2004) use bar charts to illustrate the distribution of SAH split by gender and by the eight waves of the BHPS used in the paper. In the code below this is preceded by a graph that pools the data for men across all of the waves. The second graph command uses over (wavenum) to produce the eight separate plots by wave. The figures are saved as encapsulated postscript (eps) files for subsequent use (Figures 2.1 and 2.2):

• graph bar sahex sahgood sahfair sahpoor sahvpoor if male==1 title (“bar chart for SAH, men”) ylabel (0 0.1 0.2 0.3 0.4 0.5)
• graph export “c:\stata\data\fig1.eps”, as (eps) preview (on)
• replace
• sort wavenum

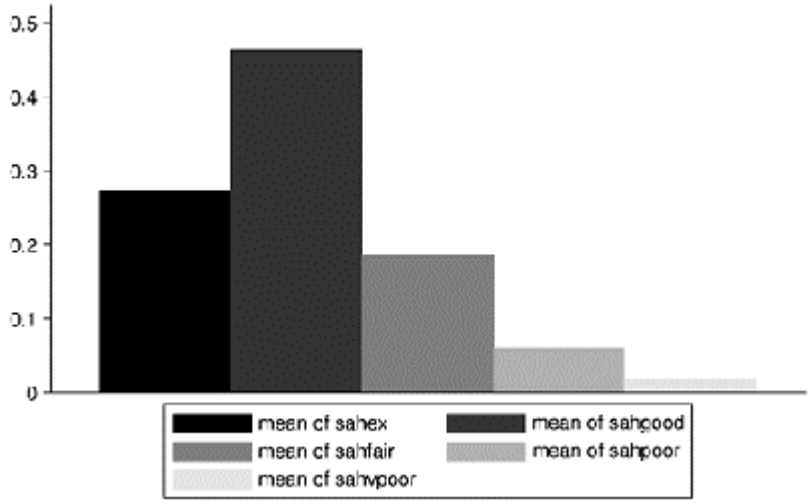

Figure 2.1 Bar chart for SAH, men.

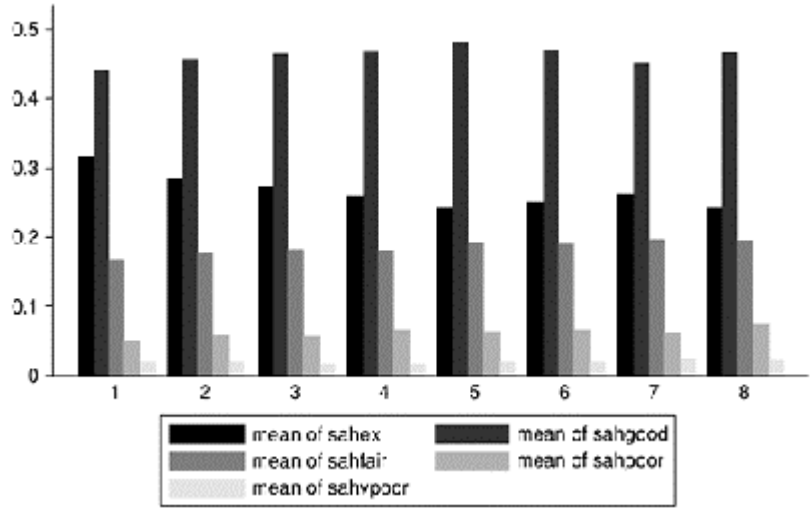

Figure 2.2 Bar chart for SAH by wave, men.

• graph bar sahex sahgood sahf air sahpoor sahvpoor if male==1, over (wavenum) title ("Bar chart for SAH by wave, men") ylabel (0 0.10.20.30.40.5)

The figures reveal the characteristic shape of the distribution of SAH. The modal category is good health, and a clear majority of respondents report either excellent or good health. The distribution is skewed, rather than symmetric, with a long right-hand tail of individuals who report fair, poor or very poor health. Comparing the distribution over time there is a decrease in the proportion reporting excellent health and an increase in those reporting fair or worse health.

The next step is to present the distribution of SAH by age group. To do this a new variable (healthtab) is created that divides individuals into ten-year age groups. The histograms are then plotted over these groups:

• gen healtab=1
  • replace healtab=2 if age<36 & age>27
  • replace healtab=3 if age<44 & age>35
  • replace healtab=4 if age<52 & age>43
  • replace healtab=5 if age<60 & age>51
  • replace healtab=6 if age<68 & age>59
  • replace healtab=7 if age<76 & age>67
  • replace healtab=8 if age<84 & age>75
  • replace healtab=9 if age>83
  • replace healtab=. if age==.
  • tab healtab
  • sort healtab

• graph bar sahex sahgood sahf air sahpoor sahvpoor if male==1 & wavenum==1, over(healtab) title("Bar chart for SAH by age group, wave1, men") ylabel (0 0.1 0.2 0.3 0.4 0.5 0.6)

The table for the new variable healthtab shows the frequency distribution across age groups:

healtab	Freq.	Percent	Cum.
1	9,612	14.85	14.85
2	10,846	16.75	31.60
3	10,121	15.63	47.23
4	10,064	15.55	62.78
5	7,270	11.23	74.01
6	6,200	9.58	83.58
7	5,842	9.02	92.61
8	3,440	5.31	97.92
9	1,346	2.08	100.00
Total	64,741	100.00	

These groups are then used in the construction of the bar chart (Figure 2.3).

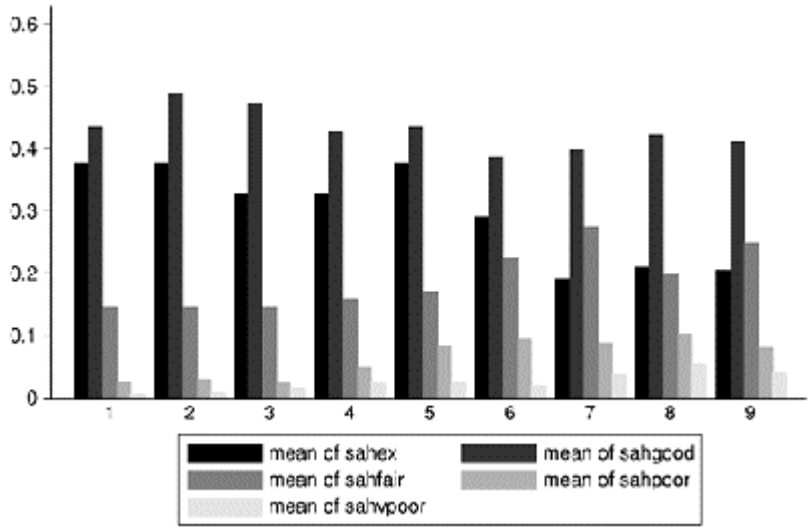

Figure 2.3 Bar chart for SAH by age group, wave 1, men.

The results help to explain the pattern observed in the previous figure. Despite the fact that respondents are asked to rate their health relative to someone of their own age, there

is a clear pattern of worsening health for the older age groups, with the proportions in the top two categories declining and the bottom three categories increasing as age increases.

To illustrate the socioeconomic gradient in SAH the distribution can be plotted for different income levels. Respondents are divided into quintiles of the distribution of income, using their average income over the panel. This can be done using the xtile command to create an indicator of the quintile that that individual belongs to (Figure 2.4):

• sort pid wavenum

• xtile incquim=meaninc if male==1,nquantiles (5)

• graph bar sahex sahgood sahfair sahpoor sahvpoor if male==1, over (incquim) ti ("Bar chart for SAH by quintile of meaninc, men") ylabel (00.10.20.30.40.5)

The figure shows that there is a clear income-related gradient in SAH. Moving from the poorest quintile (1) to the richest (5) sees an increase in the proportion reporting excellent health and a decline in the proportion reporting very poor health.

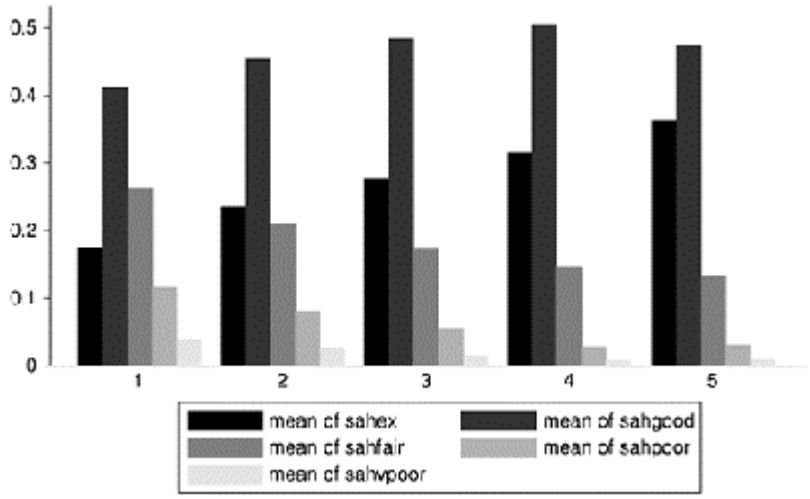

Figure 2.4 Bar chart for SAH by quintile of meaninc, men.

Another way of visualizing the income-health gradient is to plot the empirical distribution function for income, split by levels of SAH. Each of these distributions are computed separately and then plotted in the same graph (Figure 2.5).

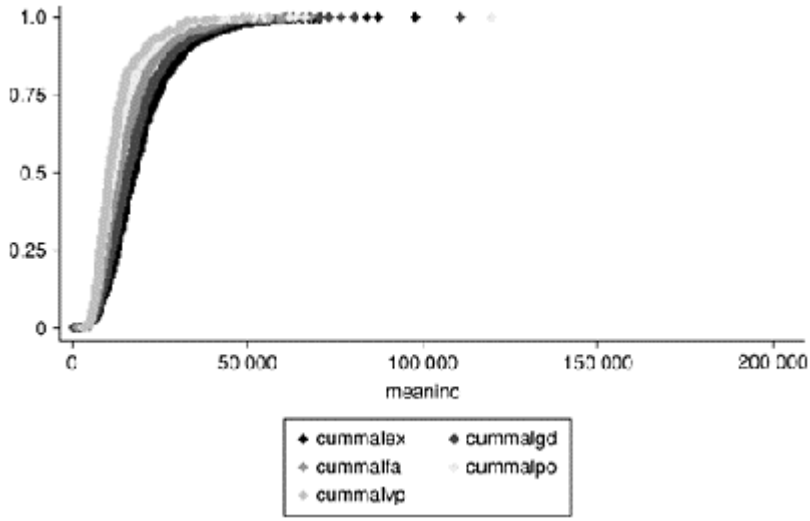

Figure 2.5 Empirical CDFs of meaninc, men.

• cumul meaninc if male==1 & sahex==1, gen (cummalex)
• cumul meaninc if male==1 & sahgood==1, gen (cummalgd)
• cumul meaninc if male==1 & sahfair==1, gen (cummalfa)
• cumul meaninc if male==1 & sahpoor==1, gen(cummalpo)
• cumul meaninc if male==1 & sahvpoor==1, gen(cummalvp)
• graph twoway scatter cummalex cummalgd cummalfa cummalpo cummalvp meaninc, s(odp.T) ylab (0(.25)1) ti ("Empirical CDF's of meaninc, men")

Moving from left to right across the graph allows a comparison of the distribution of income across increasing levels of SAH. This shows evidence of what is known as stochastic dominance: the empirical distribution functions lie to the right for those in better health.

Our second indicator of socioeconomic status is education, measured by the highest formal qualification achieved. The new variable edatt groups individuals according to increasing levels of qualification (Figure 2.6).

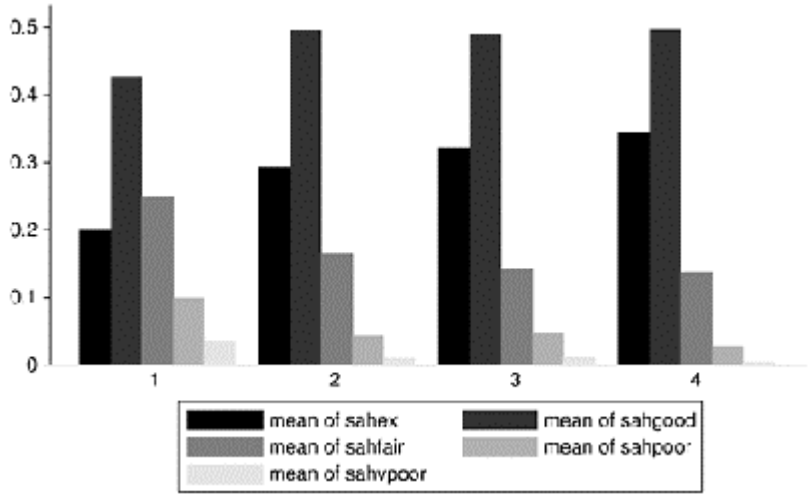

Figure 2.6 Bar chart for SAH by education, men.

sort pid wavenum
gen edatt=1
replace edatt=2 if ocse==1
replace edatt=3 if hndalev==1
replace edatt=4 if deghdeg=1
• sort edatt
• graph bar sahex sahgood sahf air sahpoor sahvpoor if male==1, over (edatt) ti (“Bar chart for SAH by education, men”) ylabel (0 0.10.20.30.40.5)

One of the aims of Contoyannis, Jones and Rice (2004) was to investigate the dynamics of health. Descriptive evidence of state dependence is provided by plotting the distribution for current SAH split by levels of SAH in the previous wave (hs tat lag) (Figure 2.7).

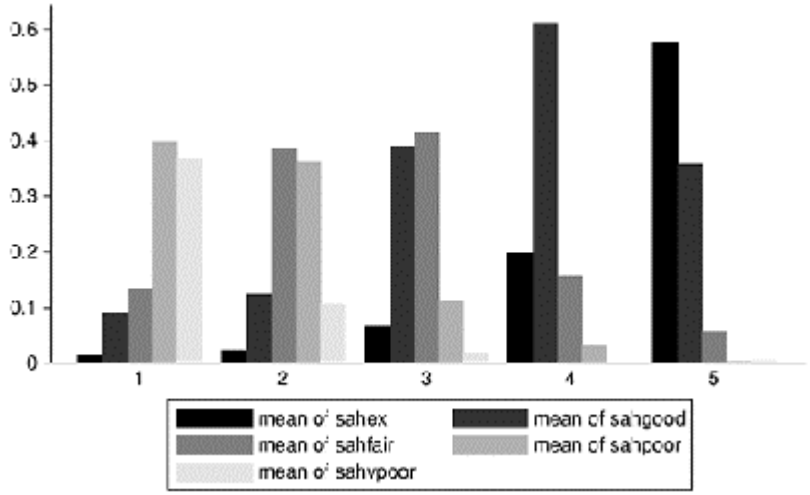

Figure 2.7 Bar chart for SAH by previous SAH, wave 2, men.

• sort hstatlag

• graph bar sahex sahgood sahfair sahpoor sahvpoor if male==1 & wavenum==2, over (hstatlag) ti ("Bar chart for SAH by previous SAH, wave 2, men") ylabel (0 0.1 0.2 0.3 0.4 0.5 0.6)

The figure reveals clear evidence of persistence in self-reported health. The probabilities of making a transition from one end of the distribution (excellent health) to the other (poor or very poor) are very small and individuals are likely to remain close to their previous level of health.

2.3 **TABLULATING THE DATA**

Along with the graphical analysis it is useful to tabulate some descriptive statistics for the data. Given the emphasis on dynamics and state dependence we begin with transition matrices. Here these are split by gender and presented for males only:

• xttrans hlstat if male==1, i (pid) t (wavenum) freq

< tr>	hlstat					
hlstat	1	2	3	4	5	Total
1	148	150	59	24	9	390
	37.95	38.46	15.13	6.15	2.31	100.00
2	152	598	473	169	37	1,429

	10.64	41.85	33.10	11.83	2.59	100.00
3	85	485	2,068	1,597	234	4,469
	1.90	10.85	46.27	35.74	5.24	100.00
4	55	251	1,696	7,402	2,069	11,473
	0.48	2.19	14.78	64.52	18.03	100.00
5	18	65	331	2,324	4,080	6,818
	0.26	0.95	4.85	34.09	59.84	100.00
Total	458	1,549	4,627	11,516	6,429	24,579
	1.86	6.30	18.83	46.85	26.16	100.00

The rows of the table indicate previous health state while the columns show current health. So, for example, the elements of the first row show the distribution of SAH at wave t, conditional on individuals having reported very poor health at wave t-1. The strong degree of persistence in SAH shows up in the high probabilities on or close to the diagonal in these tables and the low probabilities away from the diagonal.

Contoyannis, Jones and Rice (2004, Table III) show sample means of the socioeconomic variables for three different samples: using all available data for each variable, using the unbalanced sample and using the balanced sample. This gives an indication of whether the more restricted samples are comparable to the full dataset or whether there are systematic differences in terms of observable characteristics. Here the summarize command provides a range of summary statistics, not just the sample means:

- * ALL AVAILABLE DATA
  - summ $xvars

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	66323	.0881745	.2835507	0	1
nvrmar	66323	.1633672	.3697031	0	1
divsep	66323	.0682116	.2521106	0	1
deghdeg	82112	.0964536	.2952141	0	1
hndalev	82112	.2024552	.4018321	0	1
ocse	82112	.2724084	.4452016	0	1
hhsize	64741	2.788357	1.329707	1	11
nch04	64741	.1443753	.4196944	0	4
nch511	64741	.2597736	.6145583	0	6
ch1218	64741	.1833151	.4861762	0	4
age	64741	46.95723	17.77155	15	100
age2	64741	25.20804	18.17837	2.25	100
age3	64741	15.01471	15.53261	.3375	100
age4	64741	9.658935	12.80088	.050625	100
yr9293	82112	.125	.3307209	0	1
yr9394	82112	.125	.3307209	0	1
yr9495	82112	.125	.3307209	0	1

yr9596	82112	.125	.3307209	0	1
yr9697	82112	.125	.3307209	0	1
yr9798	82112	.125	.3307209	0	1
lninc	64101	9.497943	.6664307	−.1312631	13.12998

* UNBALANCED ESTIMATION SAMPLE

summ $xvars if insampm==1

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	64053	.0894103	.2853373	0	1
nvrmar	64053	.1609605	.3674973	0	1
divsep	64053	.0689585	.2533856	0	1
deghdeg	64053	.1082385	.3106838	0	1
hndalev	64053	.2152436	.4109945	0	1
ocse	64053	.2797683	.4488888	0	1
hhsize	64053	2.791204	1.329624	1	11
nch04	64053	.1450518	.4206046	0	4
nch511	64053	.2602376	.6154699	0	6
nch1218	64053	.1832701	.4859802	0	4
age	64053	46.95126	17.78103	15	100
age2	64053	25.20581	18.18994	2.25	100
age3	64053	15.01587	15.54473	.3375	100
age4	64053	9.662014	12.8131	.050625	100
yr9293	64053	.127535	.3335739	0	1
yr9394	64053	.1228982	3283229	0	1
yr9495	64053	.1163412	3206361	0	1
yr9596	64053	.1151702	3192298	0	1
yr9697	64053	.1112672	3144652	0	1
yr9798	64053	.1070207	.309142	0	1
lninc	64053	9.498008	.666476	−.1312631	13.12998

• * BALANCED ESTIMATION SAMPLE

• summ $xvars if all wave sm==1

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	48992	.079462	.2704612	0	1
nvrmar	48992	.1444113	.3515099	0	1
divsep	48992	.0676233	.2511009	0	1
deghdeg	48992	.114631	.3185793	0	1
hndalev	48992	.2261594	.4183478	0	1
ocse	48992	.2867407	.452244	0	1
hhsize	48992	2.815051	1.303281	1	10

nch04	48992	.1494121	.4218498	0	4
nch511	48992	.27133	.6221702	0	4
nch1218	48992	.186459	.4884763	0	4
age	48992	46.7817	16.98556	15	100
age2	48992	24.77031	17.23182	2.25	100
age3	48992	14.46847	14.53681	.3375	100
age4	48992	9.104977	11.8005	.050625	100
yr9293	48992	.125	.3307223	0	1
yr9394	48992	.125	.3307223	0	1
yr9495	48992	.125	.3307223	0	1
yr9596	48992	.125	.3307223	0	1
yr9697	48992	.125	.3307223	0	1
yr9798	48992	.125	.3307223	0	1
lninc	48992	9.530462	.6420103	3.324561	12.9514

The descriptive analysis is taken a stage further in Contoyannis, Jones and Rice (2004, Table IV). This compares the full sample with sub-groups who are defined according to particular sequences of reported health: those who are always in excellent or good health, those who are always in poor or very poor health, those who make a single transition away from excellent or good health (becoming unhealthy), and those who make a single transition away from poor or very poor health (becoming healthy). The following Stata code defines these groups, for the males in the sample, and runs separate summary statistics for each group:

• tab hlstat if male==1

hlstat	Freq.	Percent	Cum.
very poor	560	1.88	1.88
poor	1,838	6.16	8.03
fair	5,501	18.43	26.46
good	13,868	46.45	72.91
excellent	8,087	27.09	100.00
Total	29,854	100.00	

• gen count1=1
  • replace count1=10 if wavenum==2
  • replace count1=100 if wavenum==3
  • replace count1=1000 if wavenum==4
  • replace count1=10000 if wavenum==5
  • replace count1=100000 if wavenum==6
  • replace count1=1000000 if wavenum==7
  • replace count1=10000000 if wavenum==8

- ****always excellent/good—
  - gen hexgood=sahex==1 sahgood==1
  - gen use=hexgood*count1
  - sort pid
  - egen tot=sum (use), by (pid)
  - summ $xvars if (tot==11111111 & male==1)
  - drop use tot

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	9544	.0209556	.1432431	0	1
nvrmar	9544	.1679589	.3738494	0	1
divsep	9544	.045264	.2078936	0	1
deghdeg	9544	.1684828	.3743141	0	1
hndalev	9544	.3051132	.4604795	0	1
ocse	9544	.2816429	.4498237	0	1
hhsize	9544	2.970138	1.273814	1	10
nch04	9544	.1658634	.4420094	0	3
nch511	9544	.2825859	.637708	0	4
nch1218	9544	.2044216	.5071193	0	3
age	9544	44.22161	15.50413	15	91
age2	9544	21.95903	15.12651	2.25	82.81
age3	9544	12.01464	12.32994	.3375	75.3571
age4	9544	7.109863	9.71877	.050625	68.57496
yr9293	9544	.125	.3307362	0	1
yr9394	9544	.125	.3307362	0	1
yr9495	9544	.125	.3307362	0	1
yr9596	9544	.125	.3307362	0	1
yr9697	9544	.125	.3307362	0	1
yr9798	9544	.125	.3307362	0	1
lninc	9508	9.70625	.6180908	4.493146	12.52561

- ****always poor/very poor—
  - gen hpovpo=sahpoor==1|sahvpoor==1
  - gen use=hpovpo*count1
  - sort pid
  - egen tot=sum (use), by (pid)
  - summ $xvars if (tot==11111111 & male==1)
  - drop use tot

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	200	.06	.2380828	0	1
nvrmar	200	.03	.1710153	0	1
divsep	200	.07	.2557873	0	1
deghdeg	200	.04	.1964509	0	1

hndalev	200	.2	.4010038	0	1
ocse	200	.12	.325777	0	1
hhsize	200	2.72	1.182621	1	6
nch04	200	.04	.2422673	0	2
nch511	200	.185	.5852243	0	3
nch1218	200	.18	.4886655	0	3
age	200	53.3	11.26251	28	84
age2	200	29.671	12.67862	7.84	70.56
age3	200	17.21977	11.38129	2.1952	59.2704
ge4	200	10.40312	9.55444	.614656	49.78714
yr9293	200	.125	.3315488	0	1
yr9394	200	.125	.3315488	0	1
yr9495	200	.125	.3315488	0	1
yr9596	200	.125	.3315488	0	1
yr9697	200	.125	.3315488	0	1
yr9798	200	.125	.3315488	0	1
lninc	200	9.222452	.5673511	7.948007	10.81978

- ****single transition from excellent/good—
  - gen use=hexgood*count1
  - sort pid
  - egen tot=sum (use), by (pid)
  - summ $xvars if (tot==1 tot==11|tot==111|tot==1111 tot==11111|tot==111111|tot==1111111) & male==1
  - tab tot if (tot==1|tot==11|tot==111 tot==1111|tot== 11111|tot==111111|tot==1111111) & male==1
  - drop use tot

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	4839	.0440174	.2051549	0	1
nvrmar	4839	.2143005	.4103786	0	1
divsep	4839	.0560033	.2299519	0	1
deghdeg	4839	.1113867	.3146429	0	1
hndalev	4839	.2512916	.4338007	0	1
ocse	4839	.2335193	.4231135	0	1
hhsize	4839	2.809465	1.328786	1	11
nch04	4839	.1155197	.3815788	0	3
nch511	4839	.2140938	.5747546	0	4
nch1218	4839	.1799959	.4751414	0	3
age	4839	46.62244	18.60908	16	93
age2	4839	25.19878	19.00365	2.56	86.49
age3	4839	15.22316	16.31359	.4096	80.4357

age4	4839	9.960428	13.5342	.065536	74.8052
yr9293	4839	.1425914	.3496919	0	1
yr9394	4839	.1151064	.3191833	0	1
yr9495	4839	.0913412	.2881235	0	1
yr9596	4839	.0787353	.269353	0	1
yr9697	4839	.0673693	.2506864	0	1
yr9798	4839	.0557967	.2295523	0	1
lninc	4780	9.521568	.6978869	.0895683	11.75901

hsumi	Freq.	Percent	Cum.
1	856	17.69	17.69
11	696	14.38	32.07
111	560	11.57	43.65
1111	618	12.77	56.42
11111	479	9.90	66.32
111111	567	11.72	78.03
11111111	1,063	21.97	100.00
Total	4,839	100.00	

• ****single transition from poor/vpoor—
• gen use=hpovpo*count1
• sort pid
• egen tot=sum (use), by (pid)
• summ $xvars if (tot==1| tot==11|tot==111 | tot== 1111| tot==1111|tot==111111|tot==1111111) & male==1
• tab tot if (tot==1|tot==11|tot==111 tot==1111|tot== 11111|tot==111111|tot==1111111) & male==1

Variable	Obs	Mean	Std. Dev.	Min	Max
widowed	796	.0753769	.2641645	0	1
nvrmar	796	.1984925	.3991157	0	1
divsep	796	.1067839	.3090325	0	1
deghdeg	796	.071608	.2579999	0	1
hndalev	796	.2386935	.426553	0	1
ocse	796	.2060302	.4047067	0	1
hhsize	796	2.497487	1.329394	1	10
nch04	796	.0854271	.317599	0	2
nch511	796	.129397	.3847275	0	2
nch1218	796	.1319095	.4153494	0	3
age	796	52.13065	18.19179	16	94
age2	796	30.48131	18.69782	2.56	88.36

age3	796	19.24572	15.81715	.4096	83.0584
age4	796	12.78279	12.79799	.065536	78.0749
yr9293	796	.1344221	.3413197	0	1
yr9394	796	.1155779	.3199191	0	1
yr9495	796	.0954774	.294058	0	1
yr9596	796	.0866834	.2815475	0	1
yr9697	796	.0816583	.2740156	0	1
yr9798	796	.0778894	.268166	0	1
lninc	791	9.421001	.6465908	5.752284	11.65996

hsumi	Freq.	Percent	Cum.
1	462	58.04	58.04
11	144	18.09	76.13
111	49	6.16	82.29
1111	36	4.52	86.81
11111	25	3.14	89.95
111111	56	7.04	96.98
1111111	24	3.02	100.00
Total	796	100.00	

These tables again reveal the associations between SAH and socioeconomic characteristics. For example, those who are always in excellent or good health on average have higher incomes, are better educated and are younger than those in the other groups.

The simple statistical associations between health and socioeconomic status revealed by graphing and tabulating the data are explored in more detail in Chapters 9 and 10. These illustrate the estimation of dynamic panel data models (Chapter 9) and methods to deal with non-response (Chapter 10).

3
Inequality in health utility and self-assessed health

3.1 INTRODUCTION

In health economics, methods based on concentration curves and indices have been used for measuring inequalities and inequities in population health and health care delivery and financing (e.g., Wagstaff and van Doorslaer 2000). The health concentration curve (CC) and concentration index (CI) provide measures of relative income-related health inequality (Wagstaff, van Doorslaer and Paci 1989). Wagstaff, Paci and van Doorslaer (1991) review and compare the properties of concentration curves and indices with alternative measures of health inequality. They argue that the main advantages are that: they capture the socioeconomic dimension of health inequalities; they use information from the whole of the distribution rather than just the extremes; they give the possibility of visual representation, through the concentration curve, and allow checks of dominance relationships.

This chapter provides a further case study to illustrate descriptive analysis that includes measures of inequality and some basic cross-sectional regression models. The case study follows van Doorslaer and Jones (2003), who assess the internal validity of using the McMaster *Health Utility Index Mark III* (HUI) to scale the responses on the typical self-assessed health (SAH) question 'How do you rate your health status in general?' They compare alternative procedures to impose cardinality on the ordinal SAH responses in the context of regression analyses and decomposition of health inequality indices. The regression models they use include OLS, ordered probit and interval regression approaches. The cardinal measures of health are used to compute and to decompose concentration indices for income-related inequality in health. These results are validated by comparison with the individual variation in the 'benchmark' HUI responses obtained from the Canadian National Population Health Survey 1994–95.

As part of the in-depth component of the NPHS, respondents were asked: 'In general, how would you say your health is?' The response categories were excellent, very good, good, fair and poor. Also, each respondent was assigned an HUI score based on their response to the questions of the eight-attribute Health Utility Index Mark III health status classification system. The HUI assigns a single numerical value, between zero and one, for all possible combinations of levels of the eight self-reported health attributes. A score of one indicates perfect health.

3.2 DISTRIBUTIONAL ANALYSIS

First the data file for the NPHS needs to be opened and a log file created for the results. The syntax for these operations was shown in Chapter 2 and will not be repeated here. As a prelude to using regression models, the empirical distribution function (EDF) for HUI can be plotted using the variable hui, first for the full sample and then splitting the sample into income quintiles. This uses the cumul command to generate cumulative frequencies for hui (these are saved as chui) which are then plotted using graph (Figure 3.1):

• cumul hui, gen (chui)
  • graph twoway scatter chui hui, ti ("Empirical CDF of HUI")

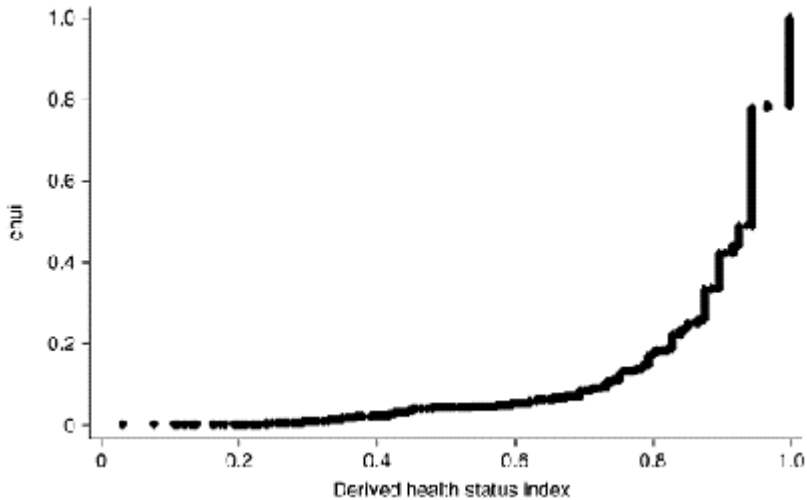

Figure 3.1 Empirical distribution function (EDF) of HUI.

The inverted 'L' shape of the EDF shows that there is a long left-hand tail made up of relatively few individuals who have low HUI scores and that many people are concentrated in the right-hand tail, with a sizeable proportion in 'full health' (HUI=1).

Next an indicator of income quintiles is generated, based on each individual's rank in the distribution of log(income) (denoted lincome in the data). This is a (less elegant) alternative to the use of the xtile command that was demonstrated in Chapter 2:

• egen rlincome=rank(lincome)
  • gen iquin=0
  • replace iquin=1 if rlincome<0.2
  • replace iquin=2 if rlincome>=0.2 & rlincome<0.4
  • replace iquin=3 if rlincome>=0.4 & rlincome<0.6
  • replace iquin=4 if rlincome>=0.6 & rlincome<0.8
  • replace iquin=5 if rlincome>=0.8

• tab iquin [aweight=nmweight]
• cumul hui if iquin==1, gen(chui1)
• cumul hui if iquin==2, gen (chui2)
• cumul hui if iquin==3, gen(chui3)
• cumul hui if iquin==4, gen (chui4)
• cumul hui if iquin==5, gen(chui5)
• graph twoway scatter chui1 chui2 chui3 chui4 chui5 hui, ti ("Empirical CDFs of HUI, by income quintile")

See Figure 3.2. This shows the income-health gradient. The EDF shifts to the right as income levels increase, implying that those in higher income groups are more concentrated in higher levels of HUI.

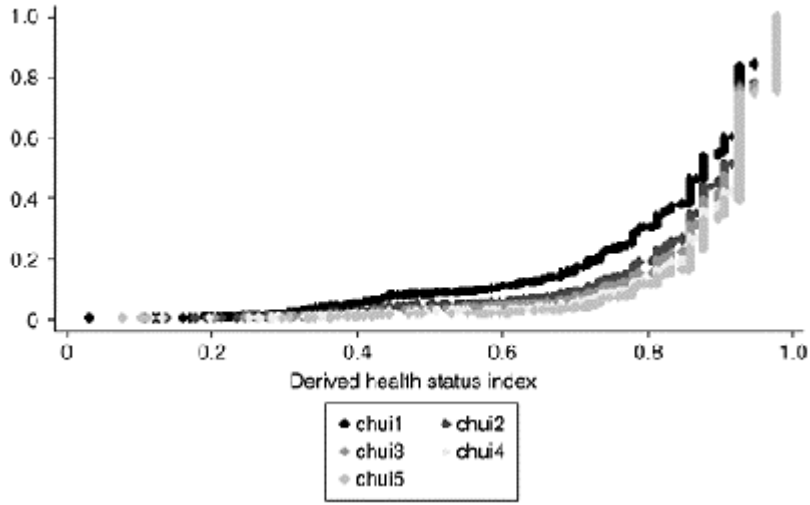

Figure 3.2 Empirical CDFs of HUI, by income quintile.

The interval regression analysis, reported below, uses cut-points based on the percentiles of the distribution of HUI that correspond to the observed cumulative probabilities of reporting each category of self-assessed health. The cumulative frequencies are obtained from the final column when sah is tabulated:

• tab sah [aweight=nmweight]

sah	Freq.	Percent	Cum.
1	373.961188	2.41	2.41
2	1,342.9465	8.64	11.05
3	4,197.96	27.01	38.06

4	5,771.6542	37.14	75.20
5	3,853.4782	24.80	100.00
Total	15,540	100.00	

Then the centile command gives the corresponding percentiles from the empirical distribution of HUI. The values 2.4, 11, 38.1 and 75.2 are the cumulative percentages from the previous table:

• centile hui, centile (2.4 11 38.1 75.2)

				--Binom. Interp.--	
Variable	Obs	Percentile	Centile	[95% Conf.	Interval]
hui	15540	2.4	.411984	.394	.418
		11	.746	.744	.753
		38.1	.897	.897	.897
		75.2	.947	.947	.947

These percentiles can be saved as scalars for future use. Notice that this command allows the data to be weighted, using the survey weights provided with the NPHS, and gives slightly different results from the previous command:

• _pctile hui [pweight=nmweight], percentiles (2.4, 11, 38.1, 75.2)
  • return list

scalars:

r(r1)=.4280000030994415
r(r2)=.7559999823570252
r(r3)=.8970000147819519
r(r4)=.9470000267028809

A detailed summary of HUI reinforces the fact that the distribution is heavily skewed, with a sizeable proportion of respondents with a value of 1 and a long left-hand tail:

• summ hui, detail

derived health status index				
	Percentiles	Smallest		
1%	.325	.031		
5%	.598	.077		
10%	.736	.077	Obs	15540
25%	.868	.104	Sum of Wgt.	15540

50%	.947		Mean	.8851377
		Largest	Std. Dev.	.1397112
75%	.947	1		
90%	1	1	Variance	.0195192
95%	1	1	Skewness	−2.39984
99%	1	1	Kurtosis	9.563778

Notice the values of the skewness and kurtosis statistics. These are quite different from those that would be expected for a normal variate: 0 and 3 respectively. They show negative skewness (long left-hand tail) and a higher level of kurtosis. A formal test for normality (sktest) is applied to HUI, using the survey weights provided. This reports the p-values and shows that the test strongly rejects the null of normality:

• sktest hui [aweight=nmweight]

Skewness/Kurtosis tests for Normality

			-------joint-----	
Variable	Pr (Skewness)	Pr (Kurtosis)	adj chi2(2)	Prob>chi2
hui	0.000	0.000		

Variants of the glcurve command can be used to produce generalized Lorenz, Lorenz, generalized concentration curves, and concentration curves for HUI. The latter use log(income) as the ranking variable, rather than ranking by HUI itself (sortvar (lincome)):

• glcurve hui [aweight=nmweight]
  • glcurve hui [aweight=nmweight], lorenz
  • glcurve hui [aweight=nmweight], sortvar (lincome)
  • glcurve hui [aweight=nmweight], sortvar (lincome) lorenz

The command can also be used to assess Lorenz dominance, illustrated here by splitting the sample according to gender and income quintiles. The first command produces concentration curves split by gender and the second Lorenz curves split by income quintiles (by (iquin)):

• glcurve hui [aweight=nmweight], sortvar (lincome) lorenz by (sex) split ti (“Concentration curves for HUI, by gender”)
  • glcurve hui [aweight=nmweight], by (iquin) split lorenz ti (“Lorenz curves for HUI, by income quintile”)

For brevity only the latter is presented here (Figure 3.3).

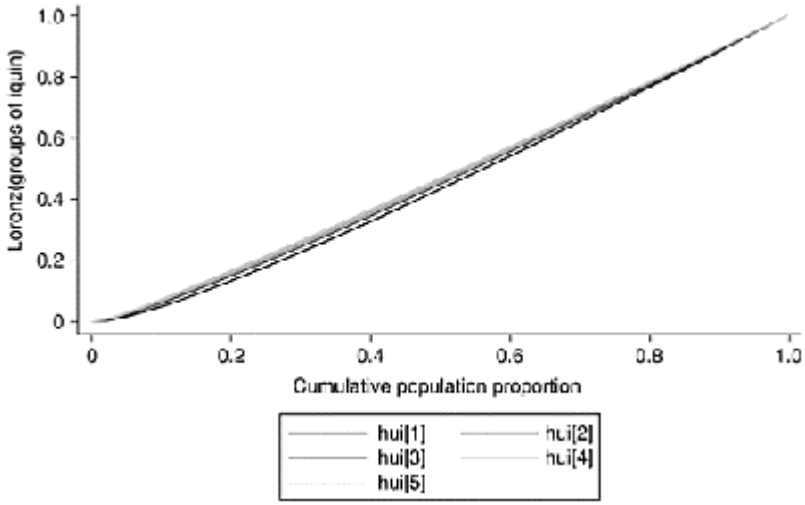

Figure 3.3 Lorenz curves for HUI, by income quintile.

3.3 REGRESSION ANALYSIS OF HUI: ORDINARY LEAST SQUARES (OLS)

Having described the distribution of HUI we now run a simple linear regression, estimated by ordinary least squares (OLS) on a set of socioeconomic characteristics that measure income (lincome), education (educ1 educ2 educ3 educ4), employment status (househ student disabled unemploy retired other), marital status (married div_wid), age and gender, which are measured here by separate age groups for men and women (m20_24 m25_29 m30_34 m35_39 m40_44 m45_49 m50_54 m55_59 m60_64 m65_69 m70_74 m75_79 m80_ f15_19 f20_24 f25_29 f30_34 f35_39 f40_44 f45_49 f50_54 f55_59 f60_64 f65_69 f70_74 f75_79 f80_):

• global xvar "lincome educ1 educ2 educ3 educ4 househ student disabled unemploy retired other married div_wid m20_24 m25_29 m30_34 m35_39 m40_44 m45_49 m50_54 m55_59 m60_64 m65_69 m70_74 m75_79 m80_ f15_19 f20_24 f25_29 f30_34 f35_39 f40_44 f45_49 f50_54 f55_59 f60_64 f65_69 f70_74 f75_79 f80_"

HUI is regressed on this list of variables, using the sample weights provided with the NPHS (Table 3.1):

• reg hui $xvar [pweight=nmweight]

Table 3.1 OLS regression for HUI

Linear regression — Number of obs= 15440; F (40,15499)= 37.04; Prob>F 0.0000; R-squared= 0.2398; Root MSE= .1156

hui	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lincome	.0095382	.0019419	4.91	0.000	.0057318	.0133445
educ1	−.0541035	.0084205	−6.43	0.000	−.0706087	−.0375982
educ2	−.0170119	.0037419	−4.55	0.000	−.0243464	−.0096773
educ3	−.0058972	.0035201	−1.68	0.094	−.012797	.0010026
educ4	−.0103114	.0027427	−3.76	0.000	−.0156875	−.0049353
househ	−.0207995	.0046358	−4.49	0.000	−.0298861	−.0117128
student	−.0020662	.00477	−0.43	0.665	−.0114159	.0072836
disabled	−.2894985	.0147468	−19.63	0.000	−.3184041	−.260593
unemploy	−.0122844	.0055427	−2.22	0.027	−.0231487	−.00142
retired	−.0331185	.006918	−4.79	0.000	−.0466786	−.0195584
other	−.0272138	.0128012	−2.13	0.034	−.0523056	−.002122
married	.008662	.0035683	2.43	0.015	.0016678	.0156562
div_wid	−.0068737	.005279	−1.30	0.193	−.0172212	.0034738
m20_24	.0093096	.0089394	1.04	0.298	−.0082126	.0268318
m25_29	−.0026133	.0092465	−0.28	0.777	−.0207374	.0155109
m30_34	−.0099412	.0100419	−0.99	0.322	−.0296244	.0097421
m35_39	−.0043818	.009723	−0.45	0.652	−.02344	.0146764
m40_44	−.0134078	.009535	−1.41	0.160	−.0320975	.0052818
m45_49	−.0235201	.0097514	−2.41	0.016	−.042634	−.0044063
m50_54	−.0306729	.0115486	−2.66	0.008	−.0533096	−.0080363
m55_59	−.0363422	.0109308	−3.32	0.001	−.0577678	−.0149166
m60_64	−.0359484	.0137488	−2.61	0.009	−.0628976	−.0089992
m65_69	−.0303558	.0129707	−2.34	0.019	−.05578	−.0049317
m70_74	−.0520681	.0163406	−3.19	0.001	−.0840976	−.0200386
m75_79	−.0438806	.0156518	−2.80	0.005	−.07456	−.0132012
m80_	−.1177298	.0215118	−5.47	0.000	−.1598955	−.0755641
f15_19	−.001686	.0098002	−0.17	0.863	−.0208955	.0175234
f20_24	−.0090165	.0096394	−0.94	0.350	−.0279109	.0098779
f25_29	−.0067698	.0096041	−0. 70	0. 481	−.025595	.0120554
f30_34	−.0069997	.0095149	−0.74	0.462	−.0256499	.0116506
f35_39	−.0120367	.0095876	−1.26	0.209	−.0308296	.0067562
f40_44	−.0220375	.0104096	−2.12	0.034	−.0424416	−.0016334

f45_49	−.0427131	.0102284	−4.18	0.000	−.062762	−.0226642
f50_54	−.047362	.0125847	−3.76	0.000	−.0720295	−.0226945
f55_59	−.0459767	.0115385	−3.98	0.000	−.0685935	−.02336
f60_64	−.0318722	.0118466	−2.69	0.007	−.0550929	−.0086515
f65_69	−.0500064	.0138387	−3.61	0.000	−.077132	−.0228809
f70_74	−.0524732	.0139375	−3.76	0.000	−.0797923	−.0251542
f75_79	−.0706702	.0156137	−4.53	0.000	−.1012748	−.0400656
f80_	−.1223062	.0177396	−6.89	0.000	−.1570778	−.0875345
cons	.8455116	.0214712	39.38	0.000	.8034256	.8875975

The residuals from this regression can be saved and summarized, with a kernel density estimate used to plot the shape of their distribution (Figure 3.4).

• predict ehui, resid
  • summehui, detail

Residuals

	Percentiles	Smallest		
1%	−.4907131	−.7663592		
5%	−.2217626	−.7468558		
10%	−.1344572	−.7092109	Obs	15540
25%	−.0373705	−.6990925	Sum of Wgt.	15540
50%	.0252128		Mean	.0006495
		Largest	Std. Dev.	.1234975
75%	.0704766	.4144598		
90%	.1017238	.4352407	Variance	.0152516
95%	.1306878	.4416342	Skewness	−1.809742
99%	.2299462	.4436584	Kurtosis	8.662831

• kdensity ehui

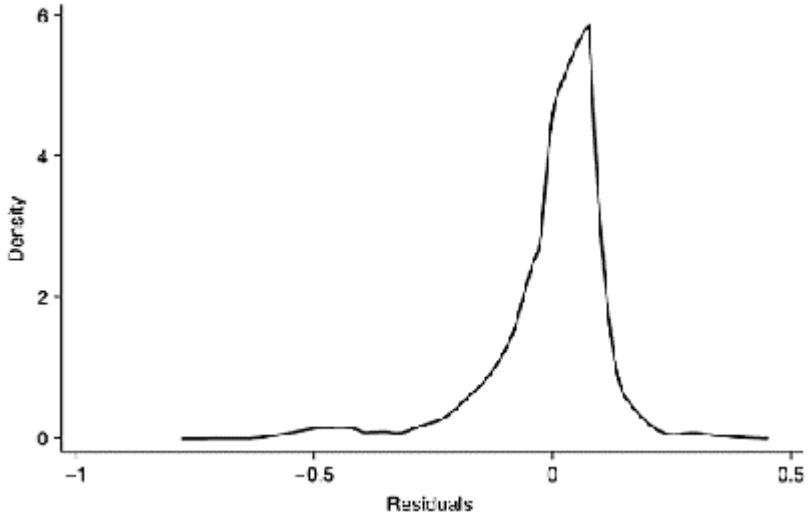

Figure 3.4 Kernel density estimates for OLS residuals.

The skewness and kurtosis statistics, along with the kernel plot, show non-normality in the distribution of the residuals (which could be confirmed by using sktest).

There is good reason to doubt the use of a simple linear regression specification with the HUI data, not least because the HUI scores are truncated at an upper limit of 1. The regression specification can be tested by a RESET test:

• predict yf
  • gen yf2=yf^2
  • quietly reg hui yf2 $xvar [pweight=nmweight]
  • test yf2

(1) yf2=0
F(1,15498)=8.76
Prob>F=0.0031

This result, coupled with the shape of the distribution of the residuals, indicates that there is a problem with mis-specification when a simple linear regression is applied to the data for HUI.

3.4 REGRESSION ANALYSIS OF SAH: ORDERED PROBIT MODEL

Self-assessed health is an ordered categorical variable and can be analysed using the ordered probit model. The ordered probit model can be used to model a discrete

dependent variable that takes ordered multinomial outcomes for each individual *i,* for example y_i=1,2……, *m*. This applies to our measure of self-assessed health (SAH), which has categorical outcomes poor, fair, good, very good and excellent. The model can be expressed as:

$$y_i = j \text{ if } \mu_{j-1} < y_i^* \leq \mu_j, \quad j = 1,....., m$$

where the latent variable, y^*, is assumed to be a linear function of a vector of socioeconomic variables x, plus a random error term ε:

$$y_i^* = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

and $\mu_0=-\infty$, $\mu_j \leq \mu_{j+1}$, $\mu_m=\infty$. Given the assumption that the error term is normally distributed, the probability of observing a particular value of y is:

$$P_{ij}=P(y_i=j)=\Phi(\mu_j-x_i\beta)-\Phi(\mu_{j-i}-x_i\beta)$$

where $\Phi(.)$ is the standard normal distribution function. If the probit link function proved to be inadequate, for example owing to the degree of skewness in the distribution, it would be possible to adopt a different distributional assumption and, hence, a different link function. With independent observations, the log-likelihood for the ordered probit model takes the form:

$$Log\ L=\Sigma_i\ \Sigma_j\ y_{ij}\ log\ P_{ij}$$

where the y_{ij} are binary variables that equal 1 if $y_i=j$. This can be maximized to give estimates of β and of the unknown threshold values μ_j.

The Stata command for the ordered probit is oprobit and predict is used to save the fitted values of the linear index (xb):

• oprobit sah $xvar [pweight=nmweight]
• predict prop, xb

Table 3.2 reports the coefficient values, which are on the latent variable scale and should not be given a quantitative interpretation. The estimates /cut1 to /cut4 are for the cut-points.

Table 3.2 Ordered probit regression for SAH

Ordered probit regression				Number of obs=	15540
				Wald chi2 (40)=	1577.31
				Prob>chi2=	0.0000
Log pseudolikelihood=−19752.678				Pseudo R2=	0.0712
sah	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]
lincome	.1649554	.0183668	8.98	0.000	.1289571 .2009536
educ1	−.5602873	.058139	−9.64	0.000	−.6742376 −.446337
educ2	−.395703	.0399178	−9.91	0.000	−.4739405 −.3174655
educ3	−.236271	.039834	−5.93	0.000	−.3143442 −.1581979
educ4	−.2110928	.0335073	−6.30	0.000	−.2767659 −.1454197
househ	−.1780776	.0390434	−4.56	0.000	−.2546011 −.101554
student	−.0426462	.0645556	−0.66	0.509	−.1691728 .0838804
disabled	−1.708831	.0824435	−20.73	0.000	−1.870417 −1.547244
unemploy	−.1130874	.0658723	−1.72	0.086	−.2421948 .01602
retired	−.3575397	.0568906	−6.28	0.000	−.4690432 −.2460362
other	−.2742358	.1045606	−2.62	0.009	−.4791708 −.0693007
married	.0391886	.0357368	1.10	0.273	−.0308543 .1092314
div_wid	.0226272	.0465103	0.49	0.627	−.0685314 .1137858
m20_24	.1417886	.141634	1.00	0.317	−.1358089 .419386
m25_29	.0442007	.1401276	0.32	0.752	−.2304443 .3188458
m30_34	−.0663295	.1409566	−0.47	0.638	−.3425994 .2099403
m35_39	−.1242918	.1395019	−0.89	0.373	−.3977105 .1491268
m40_44	−.1422974	.1411345	−1.01	0.313	−.4189159 .134321
m45_49	−.2069288	.1444126	−1.43	0.152	−.4899723 .0761146
m50_54	−.3398105	.1477348	−2.30	0.021	−.6293654 −.0502556
m55_59	−.4190716	.1478318	−2.83	0.005	−.7088166 −.1293266
m60_64	−.4244357	.1544819	−2.75	0.006	−.7272146 −.1216568
m65_69	−.4656994	.1613483	−2.89	0.004	−.7819361 −.1494626
m70_74	−.4487233	.1626481	−2.76	0.006	−.7675078 −.1299388
m75_79	−.4974953	.1672716	−2.97	0.003	−.8253417 −.169649
m80_	−.558025	.1924422	−2.90	0.004	−.9352047 −.1808453
f15_19	.0614059	.1535281	0.40	0.689	−.2395037 .3623155
f20_24	−.0097593	.141282	−0.07	0.945	−.2866669 .2671484
f25_29	−.0518077	.1422435	−0.36	0.716	−.3305999 .2269845
f30_34	−.0277856	.1394475	−0.20	0.842	−.3010977 .2455265
f35_39	−.1191202	.1411559	−0.84	0.399	−.3957807 .1575403
f40_44	−.2225667	.1418619	−1.57	0.117	−.5006108 .0554774

f45_49	−.2827923	.1452191	−1.95	0.051	−.5674165	.0018319
f50_54	−.3524339	.1457685	−2.42	0.016	−.6381349	−.0667329
f55_59	−.3924772	.1507703	−2.60	0.009	−.6879816	−.0969727
f60_64	−.3106749	.1531763	−2.03	0.043	−.6108949	−.0104549
f65_69	−.396827	.1559805	−2.54	0.011	−.7025431	−.0911108
f70_74	−.3639925	.1594664	−2.28	0.022	−.6765409	−.0514441
f75_79	−.4977896	.1633576	−3.05	0.002	−.8179647	−.1776145
f80_	−.5159627	.1714112	−3.01	0.003	−.8519224	−.1800029
/cut1	−1.19458	.2225152			−1.630701	−.758458
/cut2	−.2546723	.2223859			−.6905406	.181196
/cut3	.8076102	.2214153			.3736441	1.241576
/cut4	1.877808	.2215731			1.443533	2.312083

A RESET test can be applied to this model as well:

• predict yf, xb
  • gen yf2=yf^2
  • quietly oprobit sah yf2 $xvar [pweight=nmweight]
  • test yf2

(1) [sah]yf2=0
chi2(1)=0.03
Prob>chi2=0.8579

In this case the model passes the test.

The ordered probit model imposes the assumption of a single linear index so that the coefficients remain stable across the categories of the dependent variable. This is relaxed by the generalized ordered probit model. The model can be estimated as a whole by maximum likelihood estimation or it can be split into separate probit models (Table 3.3):

• gen cdsah1=0
  • gen cdsah2=0
  • gen cdsah3=0
  • gen cdsah4=0
  • replace cdsah1=1 if sah>l
  • replace cdsah2=1 if sah>2
  • replace cdsah3=1 if sah>3
  • replace cdsah4=1 if sah>4
  • probit cdsah1 $xvar [pweight=nmweight]
  • predict cdyf1, xb
  • probit cdsah2 $xvar [pweight=nmweight]
  • predict cdyf2, xb
  • probit cdsah3 $xvar [pweight=nmweight]
  • predict cdyf3, xb

- probit cdsah4 $xvar [pweight=nmweight]
- predict cdyf4, xb
- gen prgop=cdsah1*cdyf1+cdsah2*cdyf2+cdsah3*cdyf3+ cdsah4*cdyf4

Table 3.3 Generalized ordered probit for SAH

Probit regression					Number of obs=	15540
					Wald chi2 (40)=	.
					Prob>chi2=	.
Log pseudolikelihood=−1266.1724					Pseudo R2=	0.2819
cdsah1	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lincome	.1956211	.0448175	4.36	0.000	.1077804	.2834618
educ1	−.4244261	.116876	−3.63	0.000	−.6534989	−.1953533
educ2	−.1010504	.1095921	−0.92	0.356	−.315847	.1137462
educ3	−.1402691	.1312699	−1.07	0.285	−.3975533	.1170151
educ4	−.1208119	.1046405	−1.15	0.248	−.3259036	.0842798
househ	−.367004	.1105406	−3.32	0.001	−.5836597	−.1503484
student	−.5364086	.2202763	−2.44	0.015	−.9681423	−.104675
disabled	−1.961509	.1204865	−16.28	0.000	−2.197658	−1.72536
unemploy	−.1102092	.2941545	−0.37	0.708	−.6867415	.466323
retired	−.5909465	.1311089	−4.51	0.000	−.8479152	−.3339779
other	−1.01273	.2088029	−4.85	0.000	−1.421977	−.6034843
married	−.0346639	.119004	−0.29	0.771	−.2679074	.1985797
div_wid	−.0628016	.1317539	−0.48	0.634	−.3210344	.1954313
m20_24	−4.119277	.5327044	−7.73	0.000	−5.163359	−3.075196
m25_29	−4.71026	.5912963	−7.97	0.000	−5.86918	−3.551341
m30_34	−4.531409	.4948268	−9.16	0.000	−5.501251	−3.561566
m35_39	−4.876852	.5589403	−8.73	0.000	−5.972355	−3.781349
m40_44	−4.584727	.4784462	−9.58	0.000	−5.522465	−3.64699
m45_49	−4.772159	.4792632	−9.96	0.000	−5.711497	−3.83282
m50_54	−5.103776	.4958878	−10.29	0.000	−6.075699	−4.131854
m55_59	−5.070399	.5014349	−10.11	0.000	−6.053193	−4.087605
m60_64	−5.025081	.5006596	−10.04	0.000	−6.006356	−4.043806
m65_69	−5.299842	.4580506	−11.57	0.000	−6.197604	−4.402079
m70_74	−5.125693	.4828114	−10.62	0.000	−6.071986	−4.1794
m75_79	−5.134053	.4957245	−10.36	0.000	−6.105655	−4.162451
m80_	−5.411994	.4914733	−11.01	0.000	−6.375264	−4.448724
f15_19	−4.212844	.5528044	−7.62	0.000	−5.296321	−3.129367
f20_24	−4.055851	.5160496	−7.86	0.000	−5.06729	−3.044412
f25_29	−5.010754	.46918	−10.68	0.000	−5.930329	−4.091178
f30_34	−4.529455	.4632811	−9.78	0.000	−5.437469	−3.621441

f35_39	−4.806379	.4781316	−10.05	0.000	−5.7435	−3.869258
f40_44	−4.922795	.4749539	−10.36	0.000	−5.853688	−3.991902
f45_49	−5.107531	.5037649	−10.14	0.000	−6.094892	−4.12017
f50_54	−4.975012	.4877507	−10.20	0.000	−5.930986	−4.019039
f55_59	−5.216321	.4654698	−11.21	0.000	−6.128625	−4.304017
f60_64	−4.811212	.4793553	−10.04	0.000	−5.750731	−3.871693
f65_69	−4.978434	.4740635	−10.50	0.000	−5.907582	−4.049287
f70_74	−4.867367	.4766518	−10.21	0.000	−5.801588	−3.933147
f75_79	−5.256889	.4864836	−10.81	0.000	−6.210379	−4.303398
f80_	−5.220947	.4647546	−11.23	0.000	−6.131849	−4.310044
_cons	5.666657	.	.	.	.	.

Probit regression — Number of obs= 15540
Wald chi2 (40)= 1179.82
Prob>chi2= 0.0000
Log pseudolikelihood=−4332.0365 — Pseudo R2= 0.1978

cdsah2	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lincome	.1680696	.0269343	6.24	0.000	.1152794	.2208599
educ1	−.5654643	.0841512	−6.72	0.000	−.7303975	−.400531
educ2	−.3988678	.0707356	−5.64	0.000	−.537507	−.2602285
educ3	−.1936766	.0768494	−2.52	0.012	−.3442987	−.0430544
educ4	−.187219	.0654569	−2.86	0.004	−.3155122	−.0589259
househ	−.2818567	.0632254	−4.46	0.000	−.4057762	−.1579372
student	−.2473398	.1197708	−2.07	0.039	−.4820862	−.0125933
disabled	−1.886412	.0923838	−20.42	0.000	−2.067481	−1.705343
unemploy	−.1525503	.109632	−1.39	0.164	−.367425	.0623245
retired	−.4811156	.0781886	−6.15	0.000	−.6343624	−.3278687
other	−.3291242	.1501762	−2.19	0.028	−.6234642	−.0347842
married	.0082076	.0627244	0.13	0.896	−.1147299	.1311451
div_wid	−.146124	.0736592	−1.98	0.047	−.2904934	−.0017546
m20_24	.2814575	.3684551	0.76	0.445	−.4407013	1.003616
m25_29	.1316817	.3756804	0.35	0.726	−.6046384	.8680017
m30_34	−.1060565	.3543693	−0.30	0.765	−.8006075	.5884945
m35_39	.211661	.3599275	0.59	0.556	−.4937839	.9171058
m40_44	−.1479059	.3545863	−0.42	0.677	−.8428823	.5470705
m45_49	−.3075038	.353545	−0.87	0.384	−1.000439	.3854316
m50_54	−.5004662	.3586751	−1.40	0.163	−1.203457	.2025241
m55_59	−.5074665	.3556173	−1.43	0.154	−1.204464	.1895305
m60_64	−.4447166	.3591227	−1.24	0.216	−1.148584	.259151
m65_69	−.5312618	.3603851	−1.47	0.140	−1.237604	.17508

m70_74	−.5550044	.3638213	−1.53	0.127	−1.268081	.1580722
m75_79	−.5337466	.3679733	−1.45	0.147	−1.254961	.1874677
m80_	−.6754587	.3730616	−1.81	0.070	−1.406646	.0557286
f15_19	.2236294	.3907676	0.57	0.567	−.542261	.9895198
f20_24	.0332948	.3686971	0.09	0.928	−.6893381	.7559278
f25_29	−.1590047	.3567436	−0.45	0.656	−.8582094	.5402
f30_34	−.0367919	.3536331	−0.10	0.917	−.7299	.6563162
f35_39	−.1954803	.3512023	−0.56	0.578	−.8838241	.4928635
f40_44	−.2535474	.3536636	−0.72	0.473	−.9467154	.4396205
f45_49	−.4489704	.3542138	−1.27	0.205	−1.143217	.2452759
f50_54	−.3078843	.3554201	−0.87	0.386	−1.004495	.3887263
f55_59	−.5289404	.3533784	−1.50	0.134	−1.221549	.1636686
f60_64	−.3773868	.3562171	−1.06	0.289	−1.075559	.3207858
f65_69	−.3861096	.3669419	−1.05	0.293	−1.105302	.3330833
f70_74	−.4183939	.355983	−1.18	0.240	−1.116108	.2793198
f75_79	−.6415386	.3582466	−1.79	0.073	−1.343689	.0606118
f80_	−.6072155	.3594902	−1.69	0.091	−1.311803	.0973723
_cons	.3852014	.3756084	1.03	0.305	−.3509776	1.12138

Probit regression — Number of obs= 15540
Wald chi2 (40)= 1012.45
Prob>chi2= 0.0000
Log pseudolikelihood=−9327.5371 — Pseudo R2= 0.0965

cdsah3	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lincome	.1707651	.0231237	7.38	0.000	.1254435	.2160867
educ1	−.595921	.070228	−8.49	0.000	−.7335655	−.4582766
educ2	−.4363106	.0486549	−8.97	0.000	−.5316725	−.3409488
educ3	−.2446486	.0499675	−4.90	0.000	−.3425831	−.1467141
educ4	−.1989412	.0418159	−4.76	0.000	−.2808989	−.1169836
househ	−.1809776	.0494793	−3.66	0.000	−.2779552	−.084
student	−.0177063	.0827377	−0.21	0.831	−.1798692	.1444567
disabled	−1.458941	.1144711	−12.75	0.000	−1.6833	−1.234582
unemploy	−.2056405	.0799964	−2.57	0.010	−.3624305	−.0488505
retired	−.3167883	.0662117	−4.78	0.000	−.4465609	−.1870157
other	−.2226051	.1150966	−1.93	0.053	−.4481902	.0029801
married	.0522349	.0467155	1.12	0.264	−.0393258	.1437957
div_wid	.0569746	.0563965	1.01	0.312	−.0535604	.1675097
m20_24	.2055518	.1637372	1.26	0.209	−.1153672	.5264708
m25_29	.1985858	.1635384	1.21	0.225	−.1219436	.5191151
m30_34	−.0322601	.1633942	−0.20	0.843	−.3525068	.2879867

m35_39	−.0836703	.1635849	−0.51	0.609	−.4042908	.2369502
m40_44	−.0664008	.1659782	−0.40	0.689	−.391712	.2589104
m45_49	−.2117624	.1672221	−1.27	0.205	−.5395117	.1159869
m50_54	−.2580822	.1720884	−1.50	0.134	−.5953692	.0792048
m55_59	−.3828647	.1730742	−2.21	0.027	−.722084	−.0436454
m60_64	−.4139242	.1776578	−2.33	0.020	−.7621271	−.0657212
m65_69	−.4710132	.1813908	−2.60	0.009	−.8265326	−.1154938
m70_74	−.3693475	.1873393	−1.97	0.049	−.7365258	−.0021692
m75_79	−.5371245	.1959168	−2.74	0.006	−.9211144	−.1531345
m80_	−.4790992	.2159863	−2.22	0.027	−.9024246	−.0557738
f15_19	.1148065	.1815737	0.63	0.527	−.2410713	.4706843
f20_24	.0804107	.1641526	0.49	0.624	−.2413225	.4021439
f25_29	.0306165	.1632192	0.19	0.851	−.2892872	.3505202
f30_34	−.0426589	.1614234	−0.26	0.792	−.3590429	.2737251
f35_39	−.0815589	.1637792	−0.50	0.618	−.4025602	.2394425
f40_44	−.1698493	.1674064	−1.01	0.310	−.4979598	.1582613
f45_49	−.2310208	.1665553	−1.39	0.165	−.5574632	.0954216
f50_54	−.3402627	.1738854	−1.96	0.050	−.6810719	.0005465
f55_59	−.4347587	.1725263	−2.52	0.012	−.7729039	−.0966134
f60_64	−.2990699	.1777358	−1.68	0.092	−.6474257	.049286
f65_69	−.3837543	.1803392	−2.13	0.033	−.7372127	−.030296
f70_74	−.379399	.1809655	−2.10	0.036	−.7340849	−.0247131
f75_79	−.3881327	.1864567	−2.08	0.037	−.7535812	−.0226842
f80_	−.4637299	.1901479	−2.44	0.015	−.8364129	−.0910468
_cons	−.911769	.2785782	−3.27	0.001	−1.457772	−.3657658

Probit regression — Number of obs= 15540

Wald chi2(40)= 503.43

Prob>chi2= 0.0000

Log pseudolikelihood=−8226.1311 — Pseudo R2= 0.0549

cdsah4	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lincome	.151571	.0258095	5.87	0.000	.1009854	.2021566
educ1	−.4331911	.0820051	−5.28	0.000	−.5939182	−.272464
educ2	−.3924342	.0517096	−7.59	0.000	−.4937831	−.2910853
educ3	−.2666841	.0499563	−5.34	0.000	−.3645966	−.1687716
educ4	−.2503242	.0420153	−5.96	0.000	−.3326727	−.1679756
house	−.1332396	.0512869	−2.60	0.009	−.23376	−.0327191
student	.0134852	.0794376	0.17	0.865	−.1422096	.16918
disabled	−1.300365	.1701556	−7.64	0.000	−1.633864	−.9668664
unemploy	−.0436566	.0858084	−0.51	0.611	−.211838	.1245248

retired	−.3045645	.0808094	−3.77	0.000	−.4629481	−.1461809
other	−.193852	.1309387	−1.48	0.139	−.450487	.0627831
married	.0463186	.047813	0.97	0.333	−.0473932	.1400304
div_wid	.1046052	.0610833	1.71	0.087	−.0151158	.2243262
m20_24	.1004415	.1408707	0.71	0.476	−.17566	.3765431
m25_29	−.0461356	.1430167	−0.32	0.747	−.3264432	.2341719
m30_34	−.0735146	.1451793	−0.51	0.613	−.3580607	.2110315
m35_39	−.1937945	.1446523	−1.34	0.180	−.4773079	.0897189
m40_44	−.1998179	.1476959	−1.35	0.176	−.4892964	.0896607
m45_49	−.18177	.1511545	−1.20	0.229	−.4780275	.1144874
m50_54	−.3568497	.1553811	−2.30	0.022	−.6613911	−.0523084
m55_59	−.4185651	.1636108	−2.56	0.011	−.7392364	−.0978938
m60_64	−.4678478	.1721684	−2.72	0.007	−.8052917	−.1304039
m65_69	−.3509491	.1818959	−1.93	0.054	−.7074584	.0055603
m70_74	−.4425149	.1921343	−2.30	0.021	−.8190912	−.0659387
m75_79	−.3799316	.1971883	−1.93	0.054	−.7664136	.0065503
m80_	−.3203145	.2449146	1.31	0.191	−.8003382	.1597093
f15_19	−.0106083	.1606118	−0.07	0.947	−.3254017	.3041851
f20_24	−.0965784	.1422689	−0.68	0.497	−.3754204	.1822636
f25_29	−.0402091	.1431479	−0.28	0.779	−.320774	.2403557
f30_34	−.0097473	.1414113	−0.07	0.945	−.2869084	.2674138
f35_39	−.1214934	.1453677	−0.84	0.403	−.4064089	.1634222
f40_44	−.2501098	.1476782	−1.69	0.090	−.5395538	.0393342
f45_49	−.2451643	.1499666	−1.63	0.102	−.5390935	.0487649
f50_54	−.3942862	.15541	−2.54	0.011	−.6988841	−.0896882
f55_59	−.204146	.1592965	−1.28	0.200	−.5163615	.1080694
f60_64	−.2972272	.167496	−1.77	0.076	−.6255133	.0310589
f65_69	−.4438222	.1683445	−2.64	0.008	−.7737714	−.1138729
f70_74	−.3076041	.1909688	−1.61	0.107	−.6818962	.0666879
f75_79	−.3602391	.1889142	−1.91	0.057	−.7305042	.010026
f80_	−.3297449	.2132925	−1.55	0.122	−.7477905	.0883007
_cons	−1.756079	.2986686	−5.88	0.000	−2.341459	−1.170699

There is some evidence here that the coefficients do vary across categories. For example, the coefficients on lincome are 0.196, 0.168, 0.171 and 0.152 across the different regressions. This issue is explored in more depth in Chapter 4.

3.5 COMBINED ANALYSIS OF HUI AND SAH: INTERVAL REGRESSION

Interval, or grouped data, regression provides an alternative to the ordered probit model in cases where the values of the upper and lower limits of the intervals are known. Because the μ*'s* are known, the estimates of β are more efficient and it is possible to identify the variance of the error term σ^2 and, hence, the scale of y^* (see e.g., Jones 2000).

Van Doorslaer and Jones's (2003) approach is to use HUI scores to scale the intervals of SAH. To do this they assume that there is a stable mapping from HUI to the (latent) variable that determines reported SAH and that this applies for all individuals. This implies that an individual's rank according to HUI will correspond to their rank according to SAH and, hence, the q-th quantile of the distribution of HUI will correspond to the q-th quantile of the distribution of SAH. They adopt a non-parametric approach to estimate the thresholds (μ_j). The first step is to compute the cumulative frequency of observations for each category of SAH. Then find the quantiles of the empirical distribution function (EDF) for HUI that match these frequencies. More formally:

$$\mu^j = F^{-1}(G_j)$$

where $F^{-1}(.)$ is the inverse of the EDF of HUI and G_j is the cumulative frequency of observations for category j of SAH.

These cut-points were derived earlier and are used now to create new variables sah1 and sah2 that contain the upper and lower thresholds that correspond to each individual's reported category of self-assessed health:

• tab sah1 sah2

	sah2					
sah1	.428	.756	.897	.947	1	Total
0	450	0	0	0	0	450
.428	0	1,603	0	0	0	1,603
.756	0	0	4,132	0	0	4,132
.897	0	0	0	5,809	0	5,809
.947	0	0	0	0	3,546	3,546
Total	450	1,603	4,132	5,809	3,546	15,540

The table illustrates how individuals in the first category are allocated values between 0 and 0.428, and so on, with individuals in the top category having values between 0.947 and 1.

The mapping forms the basis for an interval regression that uses the HUI cut-points applied to the observed categories of self-assessed health. The predict command saves unconditional predictions of the linear index (xb) for the subsequent RESET test (Table 3.4):

- intreg sah1 sah2 $xvar [pweight=nmweight]
  - predict yf

Table 3.4 Interval regression for SAH

Interval regression				Number of obs=	15540
				Wald chi2 (40)=	1335.05
Log pseudolikelihood=−26239.933				Prob>chi2=	0.0000
	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf. Interval]
lincome	.015119	.0017388	8.70	0.000	.0117111 .0185269
educ1	−.0562454	.0067066	−8.39	0.000	−.0693901 −.0431007
educ2	−.0283588	.0034447	−8.23	0.000	−.0351103 −.0216073
educ3	−.0144777	.0032621	−4.44	0.000	−.0208713 −.0080841
educ4	−.0124248	.0025299	−4.91	0.000	−.0173833 −.0074664
househ	−.0146191	.003492	−4.19	0.000	−.0214633 −.0077749
student	−.0072161	.0048231	−1.50	0.135	−.0166692 .002237
disabled	−.2634327	.0162045	−16.26	0.000	−.2951928 −.2316725
unemploy	−.0072957	.0052233	−1.40	0.162	−.0175332 .0029419
retired	−.0361145	.0062233	−5.80	0.000	−.0483118 −.0239171
other	−.0349469	.0142279	−2.46	0.014	−.0628332 −.0070607
married	.0022323	.0031895	0.70	0.484	−.004019 .0084836
div_wid	−.0019504	.0046915	−0.42	0.678	−.0111455 .0072448
m20_24	.0069926	.0111023	0.63	0.529	−.0147676 .0287528
m25_29	−.0002932	.011375	−0.03	0.979	−.0225878 .0220014
m30_34	−.0093219	.0111582	−0.84	0.403	−.0311917 .0125478
m35_39	−.0111961	.0113167	−0.99	0.322	−.0333765 .0109843
m40_44	−.0129038	.0112643	−1.15	0.252	−.0349813 .0091738
m45_49	−.0204851	.0116751	−1.75	0.079	−.0433679 .0023977
m50_54	−.0313215	.0125264	−2.50	0.012	−.0558729 −.0067702
m55_59	−.0381922	.0123754	−3.09	0.002	−.0624476 −.0139369
m60_64	−.0393531	.0147708	−2.66	0.008	−.0683033 −.0104029
m65_69	−.0498255	.0159762	−3.12	0.002	−.0811382 −.0185128
m70_74	−.0440528	.015444	−2.85	0.004	−.0743224 −.0137832
m75_79	−.0515598	.0169022	−3.05	0.002	−.0846874 −.0184322
m80_	−.0678987	.0210494	−3.23	0.001	−.1091548 −.0266427
f15_19	.0034558	.0118234	0.29	0.770	−.0197175 .0266292
f20_24	−.0016696	.0112158	−0.15	0.882	−.0236522 .020313
f25_29	−.0111243	.0117724	−0.94	0.345	−.0341976 .0119491
f30_34	−.0071488	.0111614	−0.64	0.522	−.0290247 .0147271
f35_39	−.0136342	.011322	−1.20	0.229	−.0358249 .0085566

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
f40_44	−.0213613	.0115409	−1.85	0.064	−.0439811	.0012584
f45_49	−.0291626	.012312	−2.37	0.018	−.0532936	−.0050316
f50_54	−.030071	.0121509	−2.47	0.013	−.0538862	−.0062557
f55_59	−.0419838	.0130302	−3.22	0.001	−.0675225	−.0164451
f60_64	−.0265331	.0138425	−1.92	0.055	−.053664	.0005978
f65_69	−.0359813	.0140099	−2.57	0.010	−.0634403	−.0085224
f70_74	−.0332684	.0140794	−2.36	0.018	−.0608636	−.0056732
f75_79	−.0570209	.0162662	−3.51	0.000	−.0889021	−.0251397
f80_	−.0588871	.0173033	−3.40	0.001	−.092801	−.0249732
_cons	.7787193	.0198475	39.24	0.000	.7398189	.8176197
/lnsigma	−2.344311	.0175255	−133.77	0.000	−2.37866	−2.309961
sigma	.0959133	.0016809			.0926747	.0992651

The coefficients for the interval regression model are measured on the same scale as the cut-points, so they can be interpreted in terms of changes in HUI. For example, those who are out of work because of sickness (disabled), on average, report a HUI score that is 0.26 lower than those in employment (the reference category).

Because we use HUI values to scale the thresholds for SAH, the linear index for the interval regression model is measured on the same scale. The unconditional prediction of the linear index $x_i\beta$ gives us a prediction of each individual's level of HUI derived from their observed SAH. This is the level of HUI that would be predicted knowing that an individual has characteristics x. The prediction is both continuous and linear in the *x's:*

• predict predhui

• table sah, contents (n predhui mean predhui min predhui max. predhui sd predhui)

sah	N (predhui)	mean (pred~i)	min (predhui)	max (predhui)	sd (predhui)
1	450	.750351	.5373018	.934244	.1105662
2	1,603	.8168626	.5319072	.9440898	.0879004
3	4,132	.8653231	.5406754	.9513755	.0579083
4	5,809	.8849842	.5545887	.9525171	.0454433
5	3,546	.8948557	.6059171	.9477294	.0361418

• table sah, contents (p25 predhui p50 predhui p75 predhui)

sah	p25 (predhui)	med (predhui)	p75 (predhui)
1	.6253738	.7905234	.8338153
2	.7937649	.8330768	.8780249
3	.8375854	.880441	.9045454

4	.8679991	.8980366	.9138924
5	.8822401	.904195	.9188288

An alternative way of computing the predicted values from the interval regression model is to use the expected value of the linear index, conditional on the individual's observed category of SAH:

$$E(y_i^*|x_i, y_i = j) = E(y_i^*|x_i, \mu_{j-1} < y_i^* \leq \mu_j) =$$
$$x_i\beta + \sigma\{\phi(\mu_{j-1} - x_i\beta)|\sigma) - \phi(\mu_j - x_i\beta)|\sigma)\}|\{\Phi(\mu_j - x_i\beta)|\sigma)$$
$$- \Phi(\mu_{j-1} - x_i\beta)|\sigma)\}$$

This gives the level of HUI that would be predicted knowing both *x and* the category of SAH that the individual reports: knowing the category of SAH that each respondent reports provides extra information. Conditioning on this information and the way in which the individuals' characteristics, *x,* vary across categories of SAH provides a more informative set of predictions of the expected value of the underlying latent variable y*. Comparing these conditional predictions to the actual data on HUI is a useful way of assessing the predictive reliability of the interval regression method. The conditional predictions use the option e (sah1, sah2) to specify the relevant range of HUI values for each individual that correspond to their reported category of SAH:

• predict prechui, e(sah1, sah2)
  • table sah, contents (n prechui mean prechui min prechui max prechui sd prechui)

sah	N (predhui)	mean (pred~i)	min (predhui)	max (predhui)	sd (predhui)
1	450	.4013075	.3807582	.4109263	.0081019
2	1, 603	.7048874	.559067	.7353834	.0365912
3	4,132	.8401397	.8040253	.8507037	.0071204
4	5,809	.9308562	.9179217	.9338263	.0019304
5	3,546	.983282	.9807187	.9837628	.0003252

• table sah, contents (p25 prechui p50 prechui p75 prechui)

sah	p25 (predhui)	med (predhui)	p75 (predhui)
1	.3929645	.4054217	.407375
2	.7060155	.7162629	.7254909
3	.8366111	.8420505	.8450774
4	.9300864	.9314095	.9321113
5	.983166	.9833658	.9834987

3.6 OVERVIEW

The case study presented here is based on van Doorslaer and Jones (2003). Their aim is to assess the internal validity of using the McMaster *Health Utility Index Mark III* (HUI) to scale the responses on the typical self-assessed health (SAH) question 'How do you rate your health status in general?' The analysis compares alternative procedures to impose cardinality on the ordinal responses obtained. These include OLS, ordered probit and interval regression approaches. In the paper inequality and decomposition results were validated by comparison with the 'benchmark' HUI responses obtained in the Canadian National Population Health Survey 1994–95. A note of caution is that HUI in itself may underestimate the true variability in health status, since there will be additional variability within each HUI category and heterogeneity in the valuation of health states. This may be offset by measurement error in the HUI classification, which would lead to an overestimate of the true variability. The problem of measurement error in self-reported data is pursued in the next chapter.

Part II
Categorical data

4
Bias in sell-reported data

4.1 INTRODUCTION

Self-assessed health is often included in general socioeconomic surveys, such as the BHPS and the European Community Household Panel (ECHP). This kind of subjective measure of health has caused debate in the literature concerning its validity. It has been argued by some that perceived health does not correspond with actual health (see Bound 1991), while others have argued that it is a valid indicator of health (see Butler *et al.* 1987). As a self-reported subjective measure of health, SAH may be prone to measurement error. General evidence of non-random measurement error in self-reported health is reviewed in Currie and Madrian (1999) and Lindeboom (2006). Crossley and Kennedy (2002) report evidence of measurement error in a five-category SAH question. They exploit the fact that a random sub-sample of respondents to the 1995 Australian National Health Survey were asked the same version of the SAH question twice, before and after other morbidity questions. The first question was administered as part of the SF-36 questionnaire on a self-completion form, the second as part of a face-to-face interview on the main questionnaire. They found a statistically significant difference in the distribution of SAH between the two questions and evidence that these differences are related to age, income and occupation. This measurement error could be explained by a *mode of administration effect,* due to the use of self-completion and face-to-face interviews (Grootendorst *et al.* (1997) find evidence that self-completion questions reveal more morbidity); or a *framing or learning effect* by which SAH responses are influenced by the intervening morbidity questions.

It is sometimes argued that the mapping of 'true health' into SAH categories may vary with respondent characteristics. This source of measurement error has been termed 'state-dependent reporting bias' (Kerkhofs and Lindeboom 1995), 'scale of reference bias' (Groot 2000) and 'response category cut-point shift' (Sadana *et al.* 2000; Murray *et al.* 2001). This occurs if sub-groups of the population use systematically different cut-point levels when reporting their SAH, despite having the same level of 'true health'.

Regression analysis of SAH can be achieved by specifying an ordered probability model, such as the ordered probit or logit, as illustrated in Chapter 3. In the context of ordered probit models the symptoms of measurement error can be captured by making the cut-points dependent on some or all of the exogenous variables used in the model and estimating a generalized ordered probit. This requires strong *a priori* restrictions on which variables affect health and which affect reporting, in order to separately identify the influence of variables on latent health and on measurement error. It is worth noting that allowing the scaling of SAH to vary across individuals is equivalent to a

heteroskedastic specification of the underlying latent variable equation (see e.g., van Doorslaer and Jones 2003). This is because location and scale cannot be separately identified in binary and ordered choice models and, in general, it is not possible to separate measurement error from heterogeneity.

Attempts to surmount this fundamental identification problem include modelling the reporting bias using more 'objective' indicators of true health (Kerkhofs and Lindeboom 1995; Lindeboom and van Doorslaer 2004). Lindeboom and van Doorslaer (2004) analyse SAH in the Canadian National Population Health Survey and use the McMaster Health Utility Index (HUI-3) as their objective measure of health. They find evidence of cut-point shift with respect to age and gender, but not for income, education or linguistic group. Alternatively, the use of 'vignettes' has been proposed as a means of determining the cut-points independently of the health equation (King *et al.* 2004, Kapteyn *et al.* 2004).

4.2 VIGNETTES

One way of identifying individual reporting behaviour regarding health is to examine variation in the evaluation of given health states represented by hypothetical vignettes (Tandon *et al.* 2003; King *et al.* 2004; Salomon *et al.* 2004). The vignettes represent fixed levels of latent health and so all variation in the rating of them can arguably be attributed to reporting behaviour, which can be examined in relation to observed characteristics. Under the assumption that individuals rate the vignettes in the same way as they rate their own health, it is possible to identify a measure of health that is purged of reporting heterogeneity.

Murray *et al.* (2003) evaluate the vignette approach to the measurement of health, in the domain of mobility, using data from 55 countries covered by the World Health Organization Multi-Country Survey Study on Health and Responsiveness (WHO-MCS, 2000–2001). The principal objective of their analysis is to obtain comparable measures of population health that are purged of cross-country differences in the reporting of health. Reporting of health is allowed to vary with age, sex and education but there is no detailed examination of these dimensions of reporting heterogeneity or of the impact on measured health disparities. Bago d'Uva *et al.* (2006) use WHO-MCS data for China, India and Indonesia to test for systematic differences in reporting of health on six domains by sex, age, urban/rural location, education and income and to assess to what extent estimated disparities in health change when reporting differences are purged from the health measures. They find that, although homogeneous reporting by socio-demographic group is significantly rejected, the size of the reporting bias in measures of health disparities is not large. Using the vignettes method, Kapteyn *et al.* (2004) find that about half of the difference in rates of self-reported work disability between the Netherlands and the US can be attributed to reporting behaviour. Other recent surveys that include vignettes are the Survey of Health and Retirement in Europe (SHARE) and the WHO World Health Surveys, 2002–2003 (Üstün *et al.* 2003).

In this chapter we use WHO-MCS data for an Indian state (Andhra Pradesh). For illustrative purposes, we consider only one health domain, affective behaviour *(affect)*. Self-reported health in this domain is obtained from the question: 'Overall in the last 30

days, how much distress, sadness or worry did you experience?' The five response categories are: 1. Extreme, 2. Severe, 3. Moderate, 4. Mild, 5. None.

Once the data have been loaded and a log file opened, the measure of self-reported health, which is called af f in our data, is renamed with the generic label y:

• rename aff y

A random sub-sample of individuals is presented with a set of six vignettes, describing levels of distress, and asked to evaluate these hypothetical cases in the same way as they evaluate their own health for this domain (i.e., using the same five response categories and the same question except for the reference to the past 30 days). About one-quarter of the sample respond to the vignettes. As the goal of this chapter is to illustrate how vignettes can be used to identify heterogeneous reporting behaviour in self-reported health (response category cut-point shift), rather than to use all the information available, we consider only four vignettes. Extension of the procedures presented to a case with a different number of vignettes is straightforward. We use:

• Vignette 1—[Ken] remains happy and cheerful almost all the time. He is very enthusiastic and enjoys life.
• Vignette 2—[Jan] feels nervous and anxious. He is depressed nearly every day for 3–4 hours thinking negatively about the future, but feels better in the company of people or when doing something that really interests him.
• Vignette 3—[John] feels tense and on edge all the time. He is depressed nearly every day and feels hopeless. He also has low self-esteem, is unable to enjoy life, and feels that he has become a burden.
• Vignette 4—[Roberta] feels depressed all the time, weeps frequently and feels completely hopeless. She feels she has become a burden, feels it is better to be dead than alive, and often plans suicide.

The variables containing vignette ratings in the domain of *affect* are given more general names:

• rename vaff1 vig1
• rename vaff2 vig2
• rename vaff3 vig3
• rename vaff4 vig4

We test for reporting heterogeneity in relation to age (in years), sex (dummy female), education (number of schooling years, educ) and log of monthly household earnings by equivalent adult, in national currencies (lincome):

• global xvar female age educ lincome

4.3 STANDARD ORDERED PROBIT MODEL

Homogeneous reporting behaviour/no cut-point shift

We start with the standard ordered probit model (as described in Chapter 3). The assumption of homogeneous reporting behaviour that is inherent in the ordered probit model arises from the constant cut-points. If this assumption does not hold, in particular, if the cut-points vary according to some of the covariates, then imposing this restriction will lead to biased estimates of the coefficients β in the latent health index since they will reflect both health effects and reporting effects. We estimate the standard ordered probit as a baseline model in order to assess the extent to which the assumption of reporting homogeneity biases the estimated health effects. The following commands estimate the ordered probit model and save results for later comparison with the specifications that accommodate reporting heterogeneity (Table 4.1):

- oprobit y $xvar
  - mat coefs_oprob=get (_b)
  - scalar k=colsof (coefs_oprob) −4
  - mat xbs_oprob=coefs_oprob [1, 1..k]'
  - predict yf, xb

Table 4.1 Ordered probit for self-reported health *(affect)*

Ordered probit estimates				Number of obs=	5129
				LR chi2 (4)=	398.61
				Prob>chi2=	0.0000
Log likelihood=−5565.7697				Pseudo R2=	0.0346
y	Coef.	Std. Err.	Z	P>|z|	[95% Conf. Interval]
female	−.1203311	.0345834	−3.48	0.001	−.1881132 −.0525489
age	−.0147881	.0011795	−12.54	0.000	−.0170998 −.0124764
educ	.029564	.0039453	7.49	0.000	.0218315 .0372966
lincome	.0857674	.0150623	5.69	0.000	.056246 .1152889
cut1	−2.391911	.1170407	(Ancillary parameters)		
cut2	−1.417121	.1084648			
cut3	−.9561097	.1075798			
_cut4	−.2486626	.1069602			

The coefficients of the explanatory variables have a qualitative interpretation: a positive coefficient means a positive effect on the latent health index, thus a higher probability of reporting a higher category of self-reported health. The results indicate significant positive relationships between the socioeconomic variables, lincome and educ, and health, while significantly negative associations are found for female and age.

The latent health index and the coefficients are not measured in natural units. Measurement of quantitative effects of the regressors should therefore make use of marginal effects (for continuous variables) and average effects (for binary variables). We can calculate partial effects on the probabilities of reporting each health category, for each individual. Here, we illustrate with the partial effect of female on the probability of reporting the best category in the health domain *affect* (j=5, no distress, sadness or worry) and compute summary statistics:

```
• scalar mu4=_b [_cut4]
   • scalar bfemale=_b [female]
   • gen ae_p5_f emale=0
   • replace ae_p5_female=
        (1−norm(mu4−yf−bfemale))−(1−norm(mu4−yf))
                                  if !female
   • replace ae_p5_female=
        (1−norm(mu4−yf))−(1−norm(mu4−yf+bfemale))
                                  if female
   • sum ae_p5_female
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ae_p5_female	5129	−.0443943	.0042895	−.0479762	−.0260146

This shows that, on average and controlling for education, income and age, being female decreases the probability of reporting the uppermost health category by −0.044.

4.4 USING VIGNETTES TO CONTROL FOR HETEROGENEOUS REPORTING

The vignettes describe hypothetical cases and individuals are asked to rate them in the same way as they evaluate their own health. As they represent fixed levels of health, individual variation in vignette ratings must be due to reporting heterogeneity. This means that the external vignette information can be used to model the cut-points (assumed fixed in the ordered probit model) as functions of the individual characteristics. These cut-points can then be imposed on the model for self-reported health, making it possible to identify health effects (β in the latent health index) rather than a mixture of health effects and reporting effects. This can be done using the hierarchical ordered probit model (HOPIT) suggested by Tandon *et al.* (2003).

The HOPIT model has two components: the vignette component reflects reporting behaviour (that is, it models the cut-points) and the own-health component represents the relationship between the individual's own health and the observables (with cut-points determined by the vignette component). The use of vignettes to identify the cut-points relies on two assumptions. First, there must be *response consistency:* individuals classify the hypothetical cases represented by the vignettes in the same way as they rate their own health. That is, the mapping used to translate the perceived latent health of others to reported categories is the same as that governing the correspondence between own health

and reported health. The second identification assumption is *vignette equivalence:* 'the level of the variable represented by any one vignette is perceived by all respondents in the same way and on the same unidimensional scale' (King *et al.* 2004, p. 194).

The two components of the HOPIT model are linked through the cut-points, so the model does not factorize into two independent parts. The joint estimation of the two parts of the model is more efficient than a two-step procedure (Kapteyn *et al.* 2004). In this chapter we start by presenting the two-step procedure as this enables a better understanding of how the model works. Additionally, under the assumption that the covariates have the same effect on all cut-points (parallel cut-point shift) the two-step procedure can be implemented using built-in Stata commands (oprobit for the vignette model and intreg for own health). When either the parallel cut-point shift is relaxed in the two-step procedure or the one-step estimation procedure is adopted, it becomes necessary to define specific programs.

Reporting behaviour: modelling vignette ratings

We use the vignette ratings (variables vig1 to vig4) to model individual reporting behaviour. From the frequencies we can see that, despite representing fixed levels of distress, the vignette ratings show considerable variation, which can be attributed to reporting heterogeneity. This is the variation that can be exploited to test for systematic reporting heterogeneity in relation to demographic and socioeconomic characteristics and to purge health disparities across such characteristics of reporting bias:

• tab1 vig*

-> tabulation of vig1

vig1	Freq.	Percent	Cum.
1	3	0.24	0.24
2	20	1.59	1.82
3	14	1.11	2.93
4	66	5.23	8.17
5	1,158	91.83	100.00
Total	1,261	100.00	

-> tabulation of vig2

vig2	Freq.	Percent	Cum.
1	12	0.95	0.95
2	223	17.68	18.64
3	457	36.24	54.88
4	541	42.90	97.78
5	28	2.22	100.00
Total	1,261	100.00	

-> tabulation of vig3

vig3	Freq.	Percent	Cum.
1	398	31.94	31.94

2	750	60.19	92.13
3	65	5.22	97.35
4	33	2.65	100.00
Total	1,246	100.00	

-> tabulation of vig4

vig4	Freq.	Percent	Cum.
1	492	39.02	39.02
2	687	54.48	93.50
3	46	3.65	97.15
4	17	1.35	98.49
5	19	1.51	100.00
Total	1,261	100.00	

The vignette component of the HOPIT is specified in the spirit of the generalized ordered probit proposed by Terza (1985). When applied to self-reported health, this model requires that one threshold is normalized to a constant so that cut-point shift is measured relative to the baseline threshold (that is, for each covariate, what is identifiable is the difference between the impact on each cut-point and the impact on the fixed cut-point). Alternatively, identification can be achieved by assuming that each covariate can be excluded from either the cut-points or the health index (Pudney and Shields 2000). While it is difficult to maintain such assumptions in the context of self-reported health, this framework becomes more attractive when vignettes are available. Since these represent levels of health that are fixed across individuals, all systematic variation in the respective ratings can be attributed to reporting heterogeneity. In this way, the covariates are naturally excluded from the latent health index and included only in the cut-points. Despite the differences noted, we refer to the vignette component of the HOPIT as a generalized ordered probit.

Formally, let h_{ik}^{v*} be the underlying health status of vignette *k, k*=1,…, 4, perceived by individual *i*. Given that each vignette represents an exogenously determined level of health, any association between underlying health observed by individual *i,* h_{ik}^{v*}, and an individual's characteristics can be ruled out. Accordingly, $E(h_{ik}^{v*})$ is assumed to depend solely on the corresponding vignette such that:

$$h_{ik}^{v*} = \alpha_k + \varepsilon_{ik}^{v}, \qquad \varepsilon_{ik}^{v} \sim N(0,1) \tag{4.1}$$

The observed category for the vignette rating, h_{ik}^{v}, is related to H_{ik}^{v*} through the following mechanism:

$$h_{ik}^{v} = j \quad if \quad \mu_i^{j-1} \leq h_{ik}^{v*} < \mu_i^{j} \tag{4.2}$$

where $\mu_i^1 < \mu_i^2 < \ldots < \mu_i^5$ and $\mu_i^0 = -\infty$, $\mu_i^5 = \infty$, $\forall i$. The cut-points are defined as functions of covariates, *x:*

$$\mu_i^j = x_i \gamma^j \tag{4.3}$$

Note that the covariates are only included to model reporting heterogeneity in the cut-points, reflecting the assumption that all systematic variation in vignette health ratings can be attributed to reporting behaviour.

The probabilities associated with each of the 5 categories are given by:

$$\Pr(h_{ik}^{v} = j) = \Phi(x_i \gamma^j - \alpha_k) - \Phi(x_i \gamma^{j-1} - \alpha_k), \quad j = 1, \ldots, 5, \tag{4.4}$$

where Φ(.) is the cumulative standard normal distribution.

In order to model the vignette ratings, it is useful to reshape the dataset from the raw form, where each row represents one individual (identified by variable id) and there are 4 columns containing vignette ratings vig1—vig4, to a 'long form' in which the vignette ratings are contained in a single column, containing the dependent variable of the vignette model (vig). The new variable vignum represents the vignette *k, k*=1,…4, to which the observation corresponds:

• reshape long vig, i(id) j (vignum)

The output displayed after reshape describes the transformations that were applied:

(note: j=1 2 3 4)

Data	wide	->	long
Number of obs.	5129	->	20516
Number of variables	10	->	8
j variable (4 values)		->	vignum
xij variables:			
	vig1 vig2…vig4	->	vig

The parameters α_k are identified as coefficients of the dummy variables indicating to which vignette each observation corresponds (vigdum2—vigdum4, with vignette 1 as the reference category):

• tab vignum, gen (vigdum)

• drop vigdum1

Suppose one is willing to impose the restriction that the covariates affect all cut-points by the same magnitude, i.e., that there is *parallel cut-point shift*. In this case, the vignettes model can be estimated by means of a standard ordered probit for the vignette ratings with regressors vigdum2—vigdum4 and individual characteristics x_i:

• oprobit vig $xvar vigdum*

Table 4.2 contains the estimation results of the ordered probit model for vignette ratings. Before making a correspondence between these results and the generalized ordered probit specified by equations (4.1) to (4.3), it is already possible to give some interpretation regarding reporting heterogeneity. We can see, for example, that the income coefficient is significantly negative, showing that richer individuals are less likely to give positive evaluations to the vignettes. The results for females and age show the same sign as for income, but the coefficients of those variables are not statistically significant. The positive coefficient for education shows that better educated people rate the vignettes in higher categories than poorer individuals, albeit not significantly so. The influences of income and education seem to work in opposite directions.

Table 4.2 Ordered probit for vignettes ratings *(affect)*

Ordered probit estimates				Number of obs=		5029
				LRchi2 (7)=		6423.61
				Prob>chi2=		0.0000
Log likelihood=−4496.2127				Pseudo R2=		0.4167
vig	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
female	−.0257524	.0358991	−0.72	0.473	−.0961133	.0446085
age	−.0017278	.0012296	−1.41	0.160	−.0041378	.0006822
educ	.0042445	.0039818	1.07	0.286	−.0035596	.0120487
lincome	−.0381293	.0156	−2.44	0.015	−.0687047	−.0075539
vigdum2	−2.634223	.0652671	−40.36	0.000	−2.762144	−.2.506301
vigdum3	−4.580486	.0762023	−60.11	0.000	−4.729839	−.4.431132
vigdum4	−4.710569	.0764633	−61.61	0.000	−4.860434	4.560704
_cut1	−5.32186	.1327098		(Ancillary parameters)		
_cut2	−3.581968	.1279875				
_cut3	−2.807196	.1257779				
cut4	−1.560814	.1184185				

Under the hypothesis of parallel cut-point shift, the estimated coefficients of the individual characteristics enter the vectors γ_j, j=1,…, 4, with the opposite sign, that is,

e.g., the coefficient of lincome in cut-points 1 to 4 is 0.038 (z=2.44). This means that richer individuals, owing to higher standards in the domain of affective behaviour, place their cut-points higher and are thus more likely to classify a given level of distress negatively. Individual cut-points under parallel shift can be obtained from the estimates b [_cut1] to b [_cut4] and predict can be used to obtain the predicted cut-point shift by all covariates (the vignette dummies are first set equal to zero to enable the prediction of cut-point shift using predict, xb):

```
• replace vigdum2=0
    • replace vigdum3=0
    • replace vigdum4=0
    • predict minuscutptshift, xb
    • drop vigdum*
    • gen mulpar=_b [_cut1]−minuscutptshif t
    • gen mu2par=_b [_cut2]−minuscutptshif t
    • gen mu3par=_b [_cut3]−minuscutptshif t
    • gen mu4par=_b[cut4]−minuscutptshif t
```

If reporting heterogeneity is stronger at some levels of health than others, then cut-point shift is not parallel. In order to relax the assumption of parallel cut-point shift, it is necessary to define a program for the generalized ordered probit. From here on, for simplicity, we refer to the models where the hypothesis of parallel shift is not imposed, as models with *non-parallel shift*. It should, however, be understood that this does not necessarily mean that cut-point shift is non-parallel but only that the models accommodate this feature. The program gop defines the log-likelihood in the same way that it would be defined for the ordered probit model. In the oprobit, however, the only argument (args) modelled as a function of the covariates would be the latent health linear index b, while in the generalized ordered probit model with vignettes, we instead include the covariates in the cut-points m1−m4.

```
• program define gop
        version 8.0
            args lnf b m1 m2 m3 m4
            tempvar p1 p2 p3 p4 p5
            quietly {
                gen double ‘p1’=0
                gen double ‘p2’=0
                gen double ‘p3’=0
                gen double ‘p4’=0
                gen double ‘p5’=0
                replace ‘p1’=norm (‘m1’−‘b’)
                replace ‘p2’=norm (‘m2’−‘b’)−norm (‘m1’−‘b’)
                replace ‘p3’=norm (‘m3’−‘b’)−norm (‘m2’−‘b’)
                replace ‘p4’=norm (‘m4’−‘b’)−norm (‘m3’−‘b’)
                replace ‘p5’=1−norm (‘m4’−‘b’)
                replace ‘lnf’=(vig==1)*ln (‘pi’)+(vig==2)*
```

```
                                    ln('p2')
        +(vig==3)*ln ('p3')+(vig==4)*ln ('p4')+(vig==5)
                                    *ln ('p5')
    }
end
```

We need to create dummy variables for vignettes 2 to 4 again:

• tab vignum, gen (vigdum)
  • drop vigdum1

Program gop is called to estimate the generalized ordered probit model for vignette ratings with vignette dummies in the latent health index (xb) and covariates in the cut-points (mu3 to mu4):

• set matsize 50
  • ml model If gop (xb: vigdum*, nocons)
  mu1: $xvar) (mu2: $xvar) (mu3: $xvar) (mu4: $xvar)
                                    if vig!=.
  • m1 search
  • m1 maximize
  • drop vigdum*

Table 4.3 shows the results for the generalized ordered probit model for vignette ratings. Higher standards or expectations are represented by positive shifts in the cut-points. For example, age has positive coefficients across all cut-points, so older individuals have higher health standards regarding *affect,* i.e., lower probabilities of considering a given situation as corresponding to a low level of distress, albeit not significantly so. It is noticeable that the coefficients vary considerably across cut-points, which was ruled out in the model with parallel shift. In the case of female and lincome, the effects are not even monotonic, although they are mostly positive. Significant positive cut-point shift is found for lincome (mu3) and for female (mu4), meaning that individuals with higher incomes are less likely to rate a given vignette as corresponding to mild or no distress and females are less likely to consider that a given vignette corresponds to no distress.

Table 4.3 Generalized ordered probit for vignette ratings *(affect)*

					Number of obs=	5029
					Wald chi2 (3)=	4229.88
Log likelihood=−4482.0651					Prob>chi2=	0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
vigdum2	−2.640425	.0657505	−40.16	0.000	−2.769293	−2.511556
vigdum3	−4.587628	.0765945	−59.89	0.000	−4.73775	−4.437505

	vigdum4	−4.716323	.076823	−61.39	0.000	−4.866894	−4.565753
mu1							
	female	−.0183473	.0525798	−0.35	0.727	−.1214019	.0847073
	age	.0007575	.0018034	0.42	0.674	−.002777	.004292
	educ	−.0008478	.0058225	−0.15	0.884	−.0122596	.010564
	lincome	.0460535	.0229401	2.01	0.045	.0010918	.0910152
	_cons	−5.327288	.1759642	−30.27	0.000	−5.672172	−4.982405
mu2							
	female	.020826	.054476	0.38	0.702	−.085945	.1275971
	age	.0019042	.0018848	1.01	0.312	−.00179	.0055984
	educ	−.0086054	.0059776	−1.44	0.150	−.0203212	.0031105
	lincome	.0211019	.0233799	0.90	0.367	−.0247219	.0669256
	_cons	−3.470971	.1744491	−19.90	0.000	−3.812885	−3.129057
		Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mu3							
	female	−.0051097	.0582521	−0.09	0.930	−.1192817	.1090623
	age	.0035332	.0020101	1.76	0.079	−.0004065	.0074728
	educ	−.0048636	.0064316	−0.76	0.450	−.0174693	.0077422
	lincome	.0910846	.0246088	3.70	0.000	.0428522	.139317
	_cons	−3.194383	.1830168	−17.45	0.000	−3.553089	−2.835676
mu4							
	female	.1675965	.0739336	2.27	0.023	.0226894	.3125037
	age	.0019589	.0025009	0.78	0.433	−.0029429	.0068606
	educ	−.0046346	.008187	−0.57	0.571	−.0206807	.0114116
	lincome	−.0166503	.0326254	−0.51	0.610	−.0805949	.0472943
	_cons	−1.308833	.2268656	−5.77	0.000	−1.753481	−.8641847

We can test for reporting homogeneity across all cut-points for all covariates by means of a test of joint significance of all variables in all cut-points. The null hypothesis of homogeneity is strongly rejected:

• test (mu1]) ([mu2]) ([mu3]) ([mu4])

(1) [mu1] female=0
(2) [mu1] age=0
(3) [mu1] educ=0
(4) [mu1] lincome=0
(5) [mu2] female=0
(6) [mu2] age=0
(7) [mu2] educ=0
(8) [mu2] lincome=0
(9) [mu3] female=0

(10) [mu3] age=0
(11) [mu3] educ=0
(12) [mu3] lincome=0
(13) [mu4] female=0
(14) [mu4] age=0
(15) [mu4] educ=0
(16) [mu4] lincome=0
chi2(16)=38.73
Prob>chi2=0.0012

Tests of significance of individual covariates in all cut-points give strong evidence of heterogeneity by income and result in joint insignificance of the effects of the other covariates (we have however seen above that the coefficient for female in the uppermost cut-point is individually significant):

• test [mu1] female [mu2] female [mu3] female [mu4] female

(1) [mu1] female=0
(2) [mu2] female=0
(3) [mu3] female=0
(4) [mu4] female=0
chi2 (4)=6.04
Prob>chi2=0.1965

• test [mu1] age [mu2] age [mu3] age [mu4] age

(1) [mu1] age=0
(2) [mu2] age=0
(3) [mu3] age=0
(4) [mu4] age=0
chi2 (4)=3.31
Prob>chi2=0.5078

• test [mu1] educ [mu2] educ [mu3] educ [mu4] educ

(1) [mu1] educ=0
(2) [mu2] educ=0
(3) [mu3] educ=0
(4) [mu4] educ=0
chi2 (4)=2.24
Prob>chi2=0.6917

• test [mu1] lincome [mu2] lincome [mu3] lincome [mu4] lincome

(1) [mu1] lincome=0
(2) [mu2] lincome=0

(3) [mu3] lincome=0
(4) [mu4] lincome=0
chi2 (4)=21.42
Prob>chi2=0.0003

We also test the hypotheses of parallel shift, by all covariates and by each individual variable. There is strong evidence of non-parallel shift by lincome:

• test [mu1=mu2=mu3=mu4]

(1) [mu1] female−[mu2] female=0
(2) [mu1] age−[mu2] age=0
(3) [mu1] educ−[mu2] educ=0
(4) [mu1] lincome−[mu2] lincome=0
(5) [mu1] female−[mu3] female=0
(6) [mu1] age−[mu3] age=0
(7) [mu1] educ−[mu3] educ=0
(8) [mu1] lincome−[mu3] lincome=0
(9) [mu1] female−[mu4] female=0
(10) [mu1] age−[mu4] age=0
(11) [mu1] educ−[mu4] educ=0
(12) [mu1] lincome−[mu4] lincome=0
chi2 (12)=29.52
Prob>chi2=0.0033

• test [mu1] female=[mu2] female=[mu3] female=[mu4] female

(1) [mu1] female−[mu2] female=0
(2) [mu1] female−[mu3] female=0
(3) [mu1] female−[mu4] female=0
chi2 (3)=5.54
Prob>chi2=0.1363

• test [mu1] age=[mu2] age=[mu3] age=[mu4] age

(1) [mu1] age−[mu2] age=0
(2) [mu1] age−[mu3] age=0
(3) [mu1] age−[mu4] age=0
chi2 (3)=1.27
Prob>chi2=0.7352

• test [mu1] educ=[mu2] educ=[mu3] educ=[mu4] educ

(1) [mu1] educ−[mu2] educ=0
(2) [mu1] educ−[mu3] educ=0
(3) [mu1] educ−[mu4] educ=0

```
        chi2 (3)=1.15
      Prob>chi2=0.7644
```

• test [mu1] lincome=[mu2] lincome=[mu3] lincome=[mu4] lincome

```
(1) [mu1] lincome−[mu2] lincome=0
 (2) [mu1] lincome−[mu3] lincome=0
 (3) [mu1] lincome−[mu4] lincome=0
        chi2 (3)=15.63
      Prob>chi2=0.0014
```

Prediction of individual cut-points from the generalized ordered probit is straightforward:

```
predict mu1, eq (mu1)
 predict mu2, eq (mu2)
 predict mu3, eq (mu3)
 predict mu4, eq (mu4)
```

Health equation adjusted for heterogeneous reporting behaviour (cut-point shift)

As in the ordered probit, the second component of the HOPIT defines the latent level of individual own health, h_i^{s*}, and the observation mechanism that relates this latent variable to the observed categorical variable, h_i^s. The difference is that the cut-points are no longer constant parameters but can vary across individuals, being determined by the vignette component of the model. Identification derives from the response consistency and the vignette equivalence assumptions. The possibility of fixing the cut-points leads to the specification of the model for individual own health as an interval regression. It should, however, be noted that the results should not be interpreted in exactly the same way as in a traditional interval regression. Consider the application in Chapter 3, which uses an interval regression for SAH with cut-points fixed at given HUI scores. In that case, the resulting vector β as well as the prediction of the linear index $x_i\beta$ are measured on the HUI scale. The own health component of the HOPIT cannot be interpreted in the same way because the cut-points are not measured in natural units, they are only measured up to scale and location parameters, not identified in the vignette component.

The underlying health status of individual *i* can be expressed as:

$$h_i^{s*} = z_i\beta + \varepsilon_i^s, \quad \varepsilon_i^s \mid z_i \sim N(0, \sigma^2) \tag{4.5}$$

where z_i is a vector of covariates containing a constant. Here, we consider the same covariates included above in the cut-points.

The observed categorical variable, h_i^s, relates to h_i^{s*} in the following way:

$$h_i^s = j \quad if \quad \mu_i^{j-1} \le h_i^{s*} < \mu_i^j \tag{4.6}$$

where $\mu_i^1 < \mu_i^2 < \ldots < \mu_i^5$, $\mu_i^0 = -\infty$, $\mu_i^5 = \infty$, $\forall i$, and μ_i^j are defined in the reporting behaviour model. As noted above, rather than being measured in natural units as in the standard interval regression, the cut-points μ_i^j, β and σ are measured relative to the normalization of scale and location parameters in the ordered probit for the vignettes. It is assumed that $h_{ik}^{v*} h_i^{s*}$ and are independent for all i=1,…, N and k=1,…, V.

It follows that the probabilities associated with each of the five categories are given by:

$$\Pr(h_i^s = j) = \Phi\left[\frac{(\mu_i^j - x_i\beta)}{\sigma}\right] - \Phi\left[\frac{(\mu_i^{j-1} - x_i\beta)}{\sigma}\right], \quad j = 1, \ldots, 5 \tag{4.7}$$

where Φ(.) is the cumulative standard normal distribution. Each of the response probabilities enters the log-likelihood function for the HOPIT model, which is composed as the sum of the log-likelihoods of the two components. The own-health model is linked to the vignette component through the cut-points, which are driven by the vignettes and imposed on the own-health component of the HOPIT.

Before modelling own health, with cut-points determined by the vignette component, we return to the original form of the dataset (wide format, in terms of the vignette variables):

• reshape wide vig, i (id) j (vignum)

(note: j=1 2 3 4)

Data	long	->	wide
Number of obs.	20516	->	5129
Number of variables	17	->	19
j variable (4 values)	vignum	->	(dropped)
xij variables:			
	vig	->	vig1 vig2…vig4

In order to estimate the health equation using an interval regression, we need to create variables containing individual limits for the intervals, which are obtained from the individual cut-points predicted in the reporting behaviour model (vignette component). We start with the cut-points obtained in the model with *parallel shift*. Following equation (4.6), for a given observed category j for own health, the lower limit of the interval is μ_i^{j-1} and the upper limit is μ_i^j. Upper (lower) limits for the lowest (highest) categories are set as missing values:

- gen y1par=mu1par*(y==2)+mu2par*(y==3)+mu3par*(y==4)
  +mu4par*(y==5) if y>1
  - gen y2par=mu1par*(y==1)+mu2par*(y==2)+mu3par*(y==3)
  +mu4par*(y==4) if y<5

We then estimate an interval regression for own health with cut-points adjusted by parallel shift and save some results:

- intreg y1par y2par $xvar
  - predict yf, xb
  - mat coefs_intreg_parallel=get(_b)
  - mat xbs_parallel=coefs_intreg_parallel[1,1..k]'
  - predict lnsig, eq(lnsigma)
  - scalar sigma_parallel=exp (lnsig)
  - drop lnsig

The results are shown in Table 4.4. Under the assumption of parallel cut-point shift (and the assumptions of the HOPIT model), the estimated coefficients represent health effects purged from reporting bias. These are not directly comparable with the ones presented in Table 4.1 (the model assuming reporting homogeneity) as they are measured on different scales.

Table 4.4 Interval regression for self-reported health with parallel cut-point shift *(affect)*

Interval regression				Number of obs=	5129
				LRchi2(4)=	381.45
Log likelihood=−5566.2632				Prob>chi2=	0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
female	−.1842888	.0604326	−3.05	0.002	−.3027345 −.0658431
age	−.0240915	.002072	−11.63	0.000	−.0281526 −.0200304
educ	.0473783	.0069181	6.85	0.000	.0338191 .0609376
lincome	.1877962	.0263203	7.14	0.000	.1362095 .239383
_cons	−1.127265	.1873552	−6.02	0.000	−1.494474 −.7600553
/lnsigma	.5568135	.0185775	29.97	0.000	.5204022 .5932247
sigma	1.745103	.0324197			1.682704 1.809815

Observation summary:		
	0	uncensored observations
	59	left-censored observations
	3017	right-censored observations
	2053	interval observations

The variance of the latent variable in the ordered probit is fixed to 1. Therefore, in order to make the estimates of the oprobit and the intreg comparable, we multiply the vector of

coefficients obtained in the former by the estimate for sigma in the latter (which results in the estimates that would be obtained if the variance in the ordered probit was normalized to the estimate of sigma in the interval regression). We then display the comparable vectors:

```
• mat xbs_oprob_parallel_comp=
            xbs_oprob* sigma_ parallel mat
  • mat compxbs_parallel=
            (xbs_oprob_parallel_comp, xbs_parallel)
  • mat list compxbs_parallel
                        compxbs_parallel [4, 2]
                                  y1          y1
          female −.20999008 −.18428878
             age −.02580672 −.02409151
            educ  .05159227  .04737832
         lincome  .14967301  .18779625
```

Adjusting for parallel cut-point shift decreases the coefficients of female, age and educ and increases the coefficient of lincome. The largest adjustment is observed in the lincome coefficient, the only variable for which there is evidence of (parallel) cut-point shift (Table 4.2). This example shows how failure to account for cut-point shift would lead to an underestimate of the effect of income on *affect*. The age, gender and education effects on health are not significantly corrected when parallel shift is allowed for.

As for the ordered probit model, we can calculate partial effects of the covariates on the probabilities associated with each health category. In order to correct for reporting heterogeneity, so that the partial effects reflect pure health effects, the cut-points should be fixed (for example, at the sample means). We calculate here the partial effect of female on the probability of being in the category 5 (no distress, sadness or worry), using the sample average of predicted cut-point 4, and then compute summary statistics:

```
• qui sum mu4par
  • scalar avgmu4=r (mean)
  • scalar bfemale=_b [female]
  • gen ae_p5_female=0
  • replace ae_p5_female=
        ((1−norm(avgmu4−yf−bfemale))/sigma_parallel)
        −((1−norm(avgmu4−yf))/sigma_parallel) if !female
  • replace ae_p5_female=
        ((1−norm(avgmu4−yf))/sigma_parallel)
        −((1−norm(avgmu4−yf+bfemale))/sigma_parallel)
                              if female
  • summ ae_p5_female
  • drop yf ae_p5_female
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ae_p5_female	5129	−.0342973	.0082935	−.0420701	−.0061975

As expected from the decrease in the coefficient of female when parallel cut-point shift is accounted for, the average effect of female is smaller in absolute value than the one for the ordered probit model.

The procedure required to estimate an equation for own health adjusted by *non-parallel cut-point shift* (Table 4.5) involves the same steps as for the case of parallel shift. We start by defining the interval limits implied by the reporting model with non-parallel cut-point shift (generalized ordered probit in Table 4.3):

- gen y1=mul* (y==2)+mu2* (y==3)+mu3* (y==4)+mu4* (y==5) if y>l
- gen y2=mul* (y==1)+mu2* (y==2)+mu3* (y==3)+mu4* (y==4) if y<5

The syntax used for the interval regression for own health, the comparison of the estimated coefficients with the model imposing reporting homogeneity and the partial effects of female on the probability of reporting no distress (purged from reporting bias) is the same as in the parallel shift case above:

- intreg y1 y2 $xvar
  - predict yf
  - mat coefs_intreg=get (_b)

- mat xbs_intreg=coefs_intreg [1,1..k]’
  - predict lnsig, eq (lnsigma)
  - scalar sigma_intreg=exp (lnsig)
  - drop lnsig

Table 4.5 Interval regression for self-reported health with non-parallel cut-point shift *(affect)*

Interval regression				Number of obs=	5129
				LR chi2 (4)=	339.98
Log likelihood=−5561.6263				Prob>chi2=	0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
female	−.1032351	.0607156	−1.70	0.089	−.2222355 .0157652
age	−.0238416	.0020815	−11.45	0.000	−.0279212 −.0197621
educ	.0462224	.0069496	6.65	0.000	.0326015 .0598433
lincome	.160966	.0264336	6.09	0.000	.1091571 .212775
_cons	−1.014458	.1880443	−5.39	0.000	−1.383018 −.6458982

/lnsigma	.5605615	.0185724	30.18	0.000	.5241602	.5969627
sigma	1.751656	.0325325			1.68904	1.816593

Observation summary: 0 uncensored observations
59 left-censored observations
3017 right-censored observations
2053 interval observations

• mat xbs_oprob_comp=xbs_oprob* sigma_intreg
  • mat compxbs=(xbs_oprob_comp, xbs_intreg)
  • mat list compxbs

compxbs [4, 2]

	y1	y1
female	−.2107786	−.10323514
age	−.02590363	−.02384165
educ	.051786	.0462224
lincome	.15023504	.16096604

  • qui sum mu4
  • scalar avgmu4=r (mean)
  • scalar bfemale=_b [female]
  • gen ae_p5_f emale=0
  • replace ae_p5_female=
    ((1−norm(avgmu4−yf−bfemale))/sigma_intreg)
    −((1−norm(avgmu4−yf))/sigma_intreg) if !female
  • replace ae_p5_female=
    ((1−norm(avgmu4−yf))/sigma_intreg)
    −((1−norm(avgmu4−yf+bfemale))/sigma_intreg)
    if female
  • summ ae_p5_f emale
  • drop yf ae_p5_female

Variable	Obs	Mean	Std. Dev.	Min	Max
ae_p5_female	5129	−.0193312	.0044513	−.0235015	−.0041697

As a consequence of mostly positive cut-point shift (Table 4.3), the estimated effect of l income on health is higher when cut-point shift is accounted for than in the homogeneous reporting model (see compxbs). We saw above that the hypothesis of parallel shift by income is rejected in the generalized ordered probit model (Table 4.3). Relaxing the restriction of parallel cut-point shift leads to an adjustment of the estimated income effect that is smaller than occurred when observed parallel shift was imposed (compxbs_parallel). For female, allowing for non-parallel shift uncovers a positive and significant shift for the uppermost cut-point only (Table 4.3). Correcting for this cut-point shift leads to a decrease in the effect of female on an individual's own health, that now

becomes insignificant (Table 4.5). This means that the significant female effect on health found in the homogeneous reporting model (Table 4.1) was capturing a reporting effect rather than a health effect. The tests of parallel shift in the generalized ordered probit did not provide evidence of non-parallel shift by female, but the adjustment of the health effect for parallel shift is smaller than what is obtained in the non-parallel shift model, indicating the importance of the more flexible version of cut-point shift.

One-step estimation of HOPIT model

Joint estimation of the two components in a one-step procedure is more efficient than the two-step procedure illustrated in the two previous subsections (see e.g., Kapteyn *et al.* 2004). The one-step estimation of the HOPIT model requires the definition of a specific program. The program hopit defined below specifies the joint log-likelihood of the model with cut-points determined by the vignette component. Recall that the dataset is currently in wide form in terms of the vignette ratings, so these are contained in variables vig1−vig4. The individual contribution to the log-likelihood (lnf) is composed by the sum of the log-likelihoods of the own health component (interval regression with cut-points m1−m4 and dependent variable y) and of the vignette component (generalized ordered probit for vignettes vig1−vig4, with cut-points m1−m4 and health index depending on the corresponding vignette):

```
cap program drop hopit
   program define hopit
   version 8.0

args lnf b s
   b_2 b_3 b_4
   m1 m2 m3 m4

tempvar b_1 p1_1 p2_1 p3_1 p4_1 p5_1
                  p1_2 p2_2 p3_2 p4_2 p5_2
                  p1_3 p2_3 p3_3 p4_3 p5_3
                  p1_4 p2_4 p3_4 p4_4 p5_4
                  p1 p2 p3 p4 p5

quietly {
   gen double ‘p1_1’=0
   gen double ‘p2_1’=0
   gen double ‘p3_1’=0
   gen double ‘p4_1’=0
   gen double ‘p5_1’=0
   gen double ‘p1_2’=0
   gen double ‘p2_2’=0
   gen double ‘p3_2’=0
   gen double ‘p4_2’=0
   gen double ‘p5_2’=0
```

```
    gen double ‘p1_3’=0
    gen double ‘p2_3’=0
    gen double ‘p3_3’=0
    gen double ‘p4_3’=0
    gen double ‘p5_3’=0
    gen double ‘p1_4’=0
    gen double ‘p2_4’=0
    gen double ‘p3_4’=0
    gen double ‘p4_4’=0
    gen double ‘p5_4’=0

gen double ‘p1’=0
    gen double ‘p2’=0
    gen double ‘p3’=0
    gen double ‘p4’=0
    gen double ‘p5’=0
    gen double ‘b_1’=0

replace ‘p1_1’=norm (‘m1’−‘b_1’)
    replace ‘p2_1’=norm (‘m2’−‘b_1’)−norm (‘m1’−‘b_1’)
    replace ‘p3_1’=norm (‘m3’−‘b_1’)−norm (‘m2’−‘b_1’)
    replace ‘p4_1’=norm (‘m4’−‘b_1’)−norm (‘m3’−‘b_1’)
    replace ‘p5_1’=1−norm (‘m4’−‘b_1’)

replace ‘p1_2’=norm (‘m1’−‘b_2’)
    replace ‘p2_2’=norm (‘m2’−‘b_2’)−norm (‘m1’−‘b_2’)
    replace ‘p3_2’=norm (‘m3’−‘b_2’)−norm (‘m2’−‘b_2’)
    replace ‘p4_2’=norm (‘m4’−‘b_2’)−norm (‘m3’−‘b_2’)
    replace ‘p5_2’=1−norm (‘m4’−‘b_2’)

replace ‘p1_3’ =norm (‘m1’−‘b_3’)
    replace ‘p2_3’=norm (‘m2’−‘b_3’)−norm (‘m1’−‘b_3’)
    replace ‘p3_3’=norm (‘m3’−‘b_3’)−norm (‘m2’−‘b_3’)
    replace ‘p4_3’=norm (‘m4’−‘b_3’)−norm (‘m3’−‘b_3’)
    replace ‘p5_3’=1−norm (‘m4’−‘b_3’)

replace ‘p1_4’=norm (‘m1’−‘b_4’)
    replace ‘p2_4’=norm (‘m2’−‘b_4’)−norm (‘m1’−‘b_4’)
    replace ‘p3_4’=norm (‘m3’−‘b_4’)−norm (‘m2’−‘b_4’)
    replace ‘p4_4’=norm (‘m4’−‘b_4’)−norm (‘m3’−‘b_4’)
    replace ‘p5_4’=1−norm (‘m4’−‘b_4’)

replace ‘p1’=norm ((‘m1’−‘b’)/‘s’)
    replace ‘p2’=norm ((‘m2’−‘b’))/‘s’−norm ((‘m1’−‘b’)/‘s’)
    replace ‘p3’=norm ((‘m3’−‘b’)/‘s’)−norm ((‘m2’−‘b’)/‘s’)
    replace ‘p4’=norm ((‘m4’−‘b’)/‘s’)−norm ((‘m3’−‘b’)/‘s’)
```

```
    replace ‘p5’=1−norm ((‘m4’−‘b’)/‘s’)

replace ‘lnf’=(vig1==1) *ln (‘p1_1’)+(vig1==2) *ln (‘p2_1’)
                    +(vig1==3) *ln (‘p3_1’)+(vig1==4) *ln (‘p4_1’)
                    +(vig1==5)*ln (‘p5_1’)

                +(vig2==1) *ln (‘p1_2’)+(vig2==2) *ln (‘p2_2’)
                    +(vig2==3) *ln (‘p3_2’)+(vig2=4) *ln (‘p4_2’)
                    +(vig2=5) *ln (‘p5_2’)

                +(vig3==1) *ln(‘p1_3’)+(vig3=2) *ln (‘p2_3’)
                    +(vig3==3) *ln (‘p3_3’)+(vig3=4) *ln (‘p4_3’)
                    +(vig3=5) *ln (‘p5_3’)

                +(vig4==1) *ln (‘p1_4’)+(vig4==2) *ln (‘p2_4’)
                    +(vig4==3) *ln (‘p3_4’)+(vig4==4) *ln (‘p4_4’)
                    +(vig4==5) *ln (‘p5_4’)

                +(y==1) *ln (‘p1’)+(y==2) *ln (‘p2’)
                    +(y==3) *ln (‘p3’)+(y==4) *ln(‘p4’)
                    +(y==5)*ln (‘p5’)
        }
    end
```

Estimation of the HOPIT is obtained with the syntax below, which specifies that the variables contained in xvar enter the own-health index (xb) and the cut-points (mu1−mu4). The first two arguments correspond to the interval regression for own health, with cut-points mu1−mu4 determined by the generalized ordered probit for the vignettes; the constant terms of vig2−vig4 correspond to the coefficients of vignette dummies in the latent index of the vignette model. Results are stored after the estimation:

```
• set matsize 70
    • ml model If hopit (xb: $xvar) (sig:)
            (vigdum2:) (vigdum3:) (vigdum4:)
            (mu1: $xvar) (mu2: $xvar) (mu3: $xvar) (mu4: $xvar)
    • ml search
    • ml maximize
    • mat coef s=get (_b)
    • mat xbhopit=coef s [1,..1k]’
    • scalar sigma=_b [sig:_cons]
```

The estimation results for the HOPIT are shown in Table 4.6. These are comparable to the ones obtained in Tables 4.3 and 4.5.

Table 4.6 HOPIT for self-reported health with cut-point shift *(affect)*

					Number of obs=	5129
					Number of obs=	5129
					Wald chi2 (4)=	179.52
Log likelihood=−10036.353					Prob>chi2=	0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
female	−.0644674	.0769727	−0.84	0.402	−.2153311	.0863963
age	−.0231236	.0026378	−8.77	0.000	−.0282937	−.0179535
educ	.0406463	.0088854	4.57	0.000	.0232313	.0580613
lincome	.1762156	.0336661	5.23	0.000	.1102314	.2421999
_cons	−1.145077	.2418426	−4.73	0.000	−1.61908	−.6710743
sig						
_cons	1.75405	.0474907	36.93	0.000	1.660969	1.84713
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
vigdum2						
_cons	−2.642789	.0632942	−41.75	0.000	−2.766843	−2.518734
vigdum3						
_cons	−4.588767	.075577	−60.72	0.000	−4.736895	−4.440639
vigdum4						
_cons	−4.717818	.0759614	−62.11	0.000	−4.8667	−4.568937
mu1						
female	−.0552663	.0508925	−1.09	0.278	−.1550138	.0444813
age	.0008983	.0017345	0.52	0.605	−.0025013	.004298
educ	−.0001712	.0056842	−0.03	0.976	−.0113119	.0109696
lincome	.045996	.0222179	2.07	0.038	.0024497	.0895423
_cons	−5.314993	.1716738	−30.96	0.000	−5.651467	−4.978518
mu2						
female	.0144407	.0495711	0.29	0.771	−.0827169	.1115984
age	.0008409	.0016979	0.50	0.620	−.0024869	.0041688
educ	−.0047904	.0055334	−0.87	0.387	−.0156357	.006055
lincome	.0146223	.0214493	0.68	0.495	−.0274176	.0566622
_cons	−3.406308	.1613814	−21.11	0.000	−3.72261	−3.090006
mu3						
female	.029133	.0509565	0.57	0.568	−.0707399	.129006
age	.0028442	.0017472	1.63	0.104	−.0005803	.0062686
educ	−.0032747	.0057034	−0.57	0.566	−.0144531	.0079037

	lincome	.0710442	.0219089	3.24	0.001	.0281035	.1139849
	_cons	−3.059877	.1637747	−18.68	0.000	−3.380869	−2.738884
mu4							
	female	.2252111	.0592099	3.80	0.000	.1091619	.3412603
	age	.0037321	.0020132	1.85	0.064	−.0002137	.0076779
	educ	−.0137074	.0066304	−2.07	0.039	−.0267028	−.000712
	lincome	.0132743	.0261479	0.51	0.612	−.0379745	.0645232
	_cons	−1.570272	.1866132	−8.41	0.000	−1.936027	−1.204517

Tests of significance of variables in the cut-points and of parallel cut-point shift can be performed in the same way as done for the generalized ordered probit model:

• test ([mu1]) ([mu2]) ([mu3]) ([mu4])

(1) [mu1] female=0
(2) [mu1] age=0
(3) [mu1] educ=0
(4) [mu1] lincome=0
(5) [mu2] female=0
(6) [mu2] age=0
(7) [mu2] educ=0
(8) [mu2] lincome=0
(9) [mu3] female=0
(10) [mu3] age=0
(11) [mu3] educ=0
(12) [mu3] lincome=0
(13) [mu4] female=0
(14) [mu4] age=0
(15) [mu4] educ=0
(16) [mu4] lincome=0
chi2 (16)=61.99
Prob>chi2–0.0000

• test [mu1] female [mu2] female [mu3] female [mu4] female

(1) [mu1] female=0
(2) [mu2] female=0
(3) [mu3] female=0
(4) [mu4] female=0
chi2 (4)=19.59
Prob>chi2=0.0006

• test [mu1] age [mu2] age [mu3] age [mu4] age

(1) [mu1] age=0
(2) [mu2] age=0
(3) [mu3] age=0
(4) [mu4] age=0
chi2 (4)=4.50
Prob>chi2=0.3420

• test [mu1] educ [mu2] educ [mu3] educ [mu4] educ

(1) [mu1] educ=0
(2) [mu2] educ=0
(3) [mu3] educ=0
(4) [mu4] educ=0
chi2 (4)=5.16
Prob>chi2=0.2717

• test [mu1] lincome [mu2] lincome [mu3] lincome [mu4] lincome

(1) [mu1] lincome=0
(2) [mu2] lincome=0
(3) [mu3] lincome=0
(4) [mu4] lincome=0
chi2 (4)=19.91
Prob>chi2=0.0005

• test [mu1=mu2=mu3=mu4]

(1) [mu1] female−[mu2] female=0
(2) [mu1] age−[mu2] age=0
(3) [mu1] educ−[mu2] educ=0
(4) [mu1] lincome−[mu2] lincome=0
(5) [mu1] female−[mu3] female=0
(6) [mu1] age−[mu3] age=0
(7) [mu1] educ−[mu3] educ=0
(8) [mu1] lincome−[mu3] lincome=0
(9) [mu1] female−[mu4] female=0
(10) [mu1] age−[mu4] age=0
(11) [mu1] educ−[mu4] educ=0
(12) [mu1] lincome−[mu4] lincome=0
chi2 (12)=52.85
Prob>chi2=0.0000

• test [mu1] female=[mu2] female=[mu3] female=[mu4] female

```
 (1) [mu1] female−[mu2] female=0
 (2) [mu1] female−[mu3] female=0
 (3) [mu1] female−[mu4] female=0
           chi2 (3)=19.11
       Prob>chi2=0.0003
```

• test [mu1] age=[mu2] age=[mu3] age=[mu4] age

```
 (1) [mu1] age−[mu2] age=0
 (2) [mu1] age−[mu3] age=0
 (3) [mu1] age−[mu4] age=0
           chi2 (3)=2.52
       Prob>chi2=0.4726
```

• test [mu1] educ=[mu2] educ=[mu3] educ=[mu4] educ

```
 (1) [mu1] educ−[mu2] educ=0
 (2) [mu1] educ−[mu3] educ=0
 (3) [mu1] educ−[mu4] educ=0
           chi2 (3)=4.07
       Prob>chi2=0.2539
```

• test [mu1] lincome=[mu2] lincome=[mu3] lincome=
[mu4] lincome

```
 (1) [mu1] lincome−[mu2] lincome=0
 (2) [mu1] lincome−[mu3] lincome=0
 (3) [mu1] lincome−[mu4] lincome=0
           chi2 (3)=14.14
       Prob>chi2=0.0027
```

We compare the health effects adjusted by non-parallel cut-point shift with the ones estimated by the standard homogeneous reporting model and calculate adjusted partial effects of female on the probability of having no distress in the same way as done with the intreg results in the previous sub-section:

• mat xbs_oprob_comp=xbs_oprob* sigma
  • mat compxbs=(xbs_oprob_comp, xbhopit)
  • mat list compxbs

compxbs[4, 2]

	y1	y1
female	−.21106662	−.06446736
age	−.02593902	−.02312362
educ	.05185677	.04064628
lincome	.15044033	.17621566

• predict yf, eq (xb)
  • predict mu, eq (mu4)
  • qui sum mu4
  • scalar avgmu4=r (mean)
  • scalar bfemale=_b [female]
  • gen ae_p5_f emale=0
  • replace ae_p5_female=
      ((1−norm(avgmu4−yf-bfemale))/sigma_parallel)
      −((1−norm(avgmu4−yf))/sigma_parallel) if !female
  • replace ae_p5_female=
      ((1−norm(avgmu4−yf))/sigma_parallel)
      −((1−norm(avgmu4−yf+bfemale))/sigma_parallel)
  if female
  • summ ae_p5_f emale
  • drop yf ae_p5_female

Variable	Obs	Mean	Std. Dev.	Min	Max
ae_p5_female	5129	−.0122192	.0026593	−.0147351	−.0028545

Regarding the significance of variables in individual cut-points and the effects of vignette adjustment on estimated health effects β, the results of the one-step estimation of the HOPIT model (Table 4.5) are largely in line with the ones obtained in the two-step procedure (Tables 4.3 and 4.4) and essentially the same comments apply. However, the one-step procedure provides greater evidence of cut-point shift by gender and that this is a non-parallel shift.

5
Health and lifestyles

5.1 INTRODUCTION

It is widely accepted that lifestyle has an effect on individual health and that variations in health among individuals may depend upon differences in health-related behaviours. Disparities in health, for example across socioeconomic groups, are partly explained by differences in lifestyle and living conditions, and lifestyle choices are dependent on many factors including economic circumstances.

Medical studies provide evidence of a strong relationship between physical health status and behaviours: the risk of mortality in the adult population, independently of the cause of death, increases because of harmful lifestyle choices (McGinnis and Foege 1993). Recent research also shows that modifiable behavioural risk factors, such as tobacco and alcohol consumption, are the major factors responsible for the incidence of particular causes of death (Mokdad *et al.* 2004). The epidemiological literature often investigates the impact of lifestyles on health using the so-called 'Alameda Seven'—from the Alameda County survey carried out in California in 1965—which include eating and sleeping habits, tobacco and alcohol consumption and physical activity. Individual health improves as more good health practices are undertaken and mortality rates are higher for those persons who have only a few healthy behaviours, independently of their income (Belloc and Breslow 1972; Belloc 1973).

Multivariate analysis allows a deeper investigation of the association between health-related behaviours and health, and the correlation among different lifestyles. An economic approach to the health production function has the advantage of relying on structural equations and accounting for methodological problems, such as unobservable heterogeneity, omitted variables bias and endogeneity. Contoyannis and Jones (2004) propose a model of health and lifestyle that controls for individual heterogeneity. This model is developed further in Balia and Jones (2005), and their paper is the basis for the case study presented in this chapter.

Balia and Jones (2005) propose a behavioural model for health that contains socioeconomic characteristics as well as individual health-related behaviours. The main health outcome is mortality and investments in health are assumed to be endogenous and to influence longevity. The relationship between individual socioeconomic characteristics and mortality is investigated emphasizing the role of lifestyle choices. These choices are influenced by socioeconomic characteristics but, to some extent, socioeconomic characteristics themselves have a direct effect on health outcomes, controlling for lifestyle choices. Furthermore, unobservable individual heterogeneity can influence both health outcomes and health-related behaviours.

The model assumes that individuals choose the optimal level of the demand for health given a time and budget constraint and given the trade-off with other consumption goods that enhance their utility. The individual is a rational and forward-looking economic

agent who maximizes his or her lifetime utility and knows the marginal productivity of investing in health-related behaviours as well as all the parameters of the decision-making process. Future utility at each point in time depends on the probability of surviving until the next period and on past consumption and investment decisions. The idea is that individuals face a trade-off between choices that maximize their direct satisfaction and other choices that improve health. If an individual decides to improve their health, they can reduce the consumption of goods believed to be detrimental to health, and consume more goods which have beneficial effects on health. As an example, cigarettes, alcohol and high-calorie foods produce an intrinsic pleasure, but might also have a long-term negative impact on health. Individual tastes, the rate of time preference, and expectations about the probability of survival should influence the pattern of intertemporal consumption. These elements are typically hidden to the researcher.

The behavioural model provides the motivation for an econometric model for mortality that takes the form of a recursive system of equations for mortality, self-assessed health and lifestyles. The model consists of structural form equations for mortality *(m)* and health *(h)* and reduced-form equations for lifestyles *(c):*

$$\begin{aligned} m &= \pi(c, h, x, \mu_m) \\ h &= h(c, x, \mu_h) \\ c &= f(x, \mu) \end{aligned} \qquad (5.1)$$

Where x is the vector of all observable exogenous variables in the model, and μ includes unobservable factors that influence both the individual utility function (μ_U), the health outcome (μ_h) and the risk of mortality (μ_m). Heterogeneity is modelled by assuming that the error terms are correlated.

5.2 HALS DATA AND SAMPLE

To estimate the structural model for mortality we use data from the first wave of the Health and Lifestyle Survey (HALS) from 1984–85. We do not use the second wave of the HALS (1991–92) because of attrition problems. Mortality information is provided by the follow-up deaths data that were released in May 2003 (Chapter 6 uses a subsequent follow-up from 2005, but here we use the same data as Balia and Jones (2005)). Most of the 9003 individuals interviewed in 1984 have been traced on the NHS Central Register, where all causes of death are notified. The flagging process was quite lengthy because it required several checks in order to be sure that the flagging registrations were related to the person that had been interviewed.

The variable flag code in the data file enables the current status of the respondent to be identified. Respondents can be:

–on file—currently alive and flagged on the NHS register
–not nhs regist—not currently registered with the NHS—but not known to be dead
–deceased—known to be dead, and death certificate information recorded on file
–rep dead not id—reported dead to HALS, not on NHS register (may be alive)
–embarked-abroad—identified on NHS register but currently out of country
–not yet flagged—not currently flagged for various reasons (no name etc.)

In Stata we use the command:

• tab flagcode

current flagging status Feb 03	Freq.	Percent	Cum.
on file	6,506	72.26	72.26
not nhs regist.	86	0.96	73.22
deceased	2,171	24.11	97.33
rep.dead not id	1	0.01	97.35
embarked—abroad	43	0.48	97.82
no flag yet rec.	196	2.18	100.00
Total	9,003	100.00	

97.8% of the sample had been flagged. Deaths account for some 24% of the original sample. The mortality data allows us to measure health outcomes up to 2003. The analysis covers a relatively long follow-up period, so increased risk of mortality may reflect the cumulative effect of poor health. In this framework the risk of mortality is defined as a function of observed characteristics, optimal level of investment in health-related behaviours and health at the time of the HALS. Hence this allows us to explain to what extent individual characteristics measured in 1984 determine subsequent health.

The sample

For the statistical analysis in this study, the original sample size has been reduced to 3,655 individuals according to item non-response: only individuals who answered all the questions relevant to the analysis are in the sample. In order to avoid confounding mortality with accidents, injuries or a genetic predisposition towards early death not related to lifestyle, only individuals 40 years of age and over at the time of the first survey are retained for analysis.

The main variables used from the HALS sample are health (death sah) and lifestyle (nsmoker breakfast sleepgd alqprud nobese exercise); socioeconomic indicators (sc12 sc3 sc45 lhqdg lhqhndA lhqO lhqnone lhqoth full part unemp sick retd keephse wkshft1); geographical and area indicators (wales north nwest yorks wmids emids anglia swest london scot rural suburb); marital status (married widow divorce seprd single); ethnicity (ethwheur); demographic characteristics (male height age age2 housown hou); and parental smoking and drinking behaviours (smother mothsmo fathsmo bothsmo alpa alma). These variables are listed as a global:

• global vars “death sah nsmoker breakfast sleepgd alqprud nobese exercise sc12 sc3 sc45 lhqdg lhqhndA lhqO lhqnone lhqoth married widow divorce seprd single full part unemp sick retd keephse wkshft1 wales north nwest yorks wmids emids anglia swest london scot rural suburb ethwheur male height age age2 housown hou smother mothsmo fathsmo bothsmo alpa alma”

A first simple investigation of the dataset consists of describing the information available and summarizing the variables of interest:

• describe $vars

variable name	storage type	display format	value label	variable label
death	float	%9.0g		1 if has died at May 2003, 0 alive
sah	float	%9.0g		1 if self-assessed health is excellent or good, 0 if fair or poor
nsmoker	float	%9.0g		1 if does not smoke, 0 if current smoker
breakfast	float	%9.0g		1 if does a healthy breakfast
sleepgd	float	%9.0g		1 if sleeps between 7 and 9 hours
alqprud	float	%9.0g		1 if consume alcohol prudently
nobese	float	%9.0g		1 if is not obese
exercise	float	%9.0g		1 if did physical exercise in the last fortnight
sc12	float	%9.0g		1 if professional/student or managerial /intermediate
sc3	float	%9.0g		1 if skilled or armed service
sc45	float	%9.0g		1 if partly skilled, unskilled, unclass. or never occupied
lhqdg	byte	%9.0g		1 if University degree
lhqhndA	float	%9.0g		1 if higher vocational qualifications or A level or equivalent
lhqO	byte	%9.0g		1 if 0 level/CSE
lhqnone	byte	%9.0g		1 if no qualification
lhqoth	byte	%9.0g		1 if other vocational /professional qualifications
married	byte	%8.0g		1 if married
widow	byte	%8.0g		1 if widow
divorce	byte	%8.0g		1 if divorced
seprd	byte	%8.0g		1 if separated
single	byte	%8.0g		1 if single
full	byte	%8.0g		1 if full time worker or student
part	byte	%8.0g		1 if part time worker
unemp	byte	%9.0g		1 if the individual unemployed
sick	byte	%9.0g		1 if absent from work due to sickness
retd	byte	%8.0g		1 if retired
keephse	byte	%8.0g		1 if housekeeper
wkshft1	float	%9.0g		1 if shift worker
wales	byte	%8.0g		1 if lives in Wales
north	byte	%8.0g		1 if lives in North
nwest	byte	%8.0g		1 if lives in North West
yorks	byte	%8.0g		1 if lives in Yorkshire

wmids	byte	%8.0g	1 if lives in West Midlands
emids	byte	%8.0g	1 if lives in East Midlands
anglia	byte	%8.0g	1 if lives in East Anglia
swest	byte	%8.0g	1 if lives in South West
london	byte	%8.0g	1 if lives in London
scot	byte	%8.0g	1 if lives in Scotland
rural	byte	%8.0g	1 if lives in the countryside
suburb	byte	%8.0g	1 if lives in the suburbs of the city
ethwheur	byte	%8.0g	1 if White European
male	byte	%9.0g	1 if male
height	byte	%9.0g	height in inches
age	double	%10.0g	age in years
age2	float	%9.0g	age /100
housown	byte	%9.0g	1 if own or rent house
hou	byte	%9.0g	number of other people in the house
smother	byte	%4.0g	1 if anyone else in house smoked
mothsmo	float	%9.0g	1 if only mother smoked
fathsmo	float	%9.0g	1 if only father smoked
bothsmo	float	%9.0g	1 if both parents smoked
alpa	byte	%9.0g	father, non to heavy drinker (0-4)
alma	byte	%9.0g	mother, non to heavy drinker (0-4)

• summarize $vars

Variable	Obs	Mean	Std. Dev.	Min	Max
death	3655	.3592339	.4798415	0	1
sah	3655	.7025992	.4571769	0	1
nsmoker	3655	.6995896	.4584992	0	1
breakfast	3655	.7069767	.4552113	0	1
sleepgd	3655	.5824897	.493216	0	1
alqprud	3655	.8798906	.3251339	0	1
nobese	3655	.8533516	.3538035	0	1
exercise	3655	.323119	.4677317	0	1
sc12	3655	.3154583	.4647617	0	1
sc3	3655	.4667579	.498962	0	1
sc45	3655	.2177839	.4127962	0	1
lhqdg	3655	.1250342	.3308029	0	1
lhqhndA	3655	.1247606	.3304925	0	1
lhqO	3655	.0943912	.2924123	0	1

lhqnone	3655	.6082079	.4882174	0	1
lhqoth	3655	.047606	.2129603	0	1
married	3655	.7606019	.4267745	0	1
widow	3655	.1277702	.3338794	0	1
divorce	3655	.0383037	.1919547	0	1
seprd	3655	.0166895	.1281227	0	1
single	3655	.0566347	.231175	0	1
full	3655	.3641587	.4812593	0	1
part	3655	.1318741	.3384002	0	1
unemp	3655	.0303694	.1716249	0	1
sick	3655	.0331053	.1789361	0	1
retd	3655	.3387141	.4733373	0	1
keephse	3655	.1017784	.302398	0	1
wkshft1	3655	.0574555	.2327428	0	1
wales	3655	.0577291	.2332625	0	1
north	3655	.0651163	.2467647	0	1

Variable	Obs	Mean	Std. Dev.	Min	Max
nwest	3655	.1277702	.3338794	0	1
yorks	3655	.0861833	.2806729	0	1
wmids	3655	.0801642	.2715843	0	1
emids	3655	.0766074	.2660039	0	1
anglia	3655	.0399453	.1958575	0	1
swest	3655	.0883721	.2838741	0	1
london	3655	.0943912	.2924123	0	1
scot	3655	.09658	.2954255	0	1
rural	3655	.2183311	.4131698	0	1
suburb	3655	.471409	.4992502	0	1
ethwheur	3655	.978933	.1436275	0	1
male	3655	.4552668	.4980631	0	1
height	3655	65.95021	3.703241	54	79
age	3655	57.46802	11.67334	40	96.8
age2	3655	34.38802	14.07611	16	93.7024
housown	3655	.9658003	.1817667	0	1
hou	3655	1.650889	1.27226	0	10
smother	3655	.3507524	.4772709	0	1
mothsmo	3655	.0309166	.1731153	0	1
fathsmo	3655	.5950752	.4909446	0	1
bothsmo	3655	.2456908	.430555	0	1
alpa	3655	1.891382	1.20047	0	4
alma	3655	.9119015	.9811625	0	4

• tab death

death	Freq.	Percent	Cum.
0	2,342	64.08	64.08
1	1,313	35.92	100.00
Total	3,655	100.00	

Around 70% of the individuals interviewed in 1984 reported having good or excellent health status relative to people of their own age. The sample comprises 46% men and 54% women, and is made up of individuals whose behaviours are mostly healthy. A high proportion of the sample are prudent in the consumption of alcohol (88%) and are not obese (85%). Only 30% of individuals are smokers, while 32% of them devote time to physical activities, 71% usually eat breakfast and 58% sleep a healthy number of hours. As for the socioeconomic characteristics, individuals are largely concentrated in skilled occupations (sc3) (about 47%), while only 32% of the sample belong to professional and managerial occupations (sc12) and 22% to semi- and non-skilled occupations (sc45). Around 61% of the respondents do not have formal educational qualifications, and only 13% have a university degree.

5.3 DESCRIPTIVE ANALYSIS

Lifestyle and socioeconomic status

We are interested in the response of different socioeconomic groups to risky behaviours. We define a list of six lifestyle indicators:

• global lifestyles “nsmoker breakfast sleepgd alqprud nobese exercise”

A simple way to investigate the relationship between individual lifestyle and socioeconomic characteristics consists of computing the partial correlation of each lifestyle with different social classes and levels of education. We use a foreach command to loop over each lifestyle while executing the command pcorr. The command pcorr allows calculation of the partial correlation coefficients of lifestyles with the top and bottom social class, holding sc3 constant:

• foreach x of global lifestyles!
pcorr ‘x’ sc12 sc45
}

Partial correlation of nsmoker with

Variable	Corr.	Sig.
sc12	0.1006	0.000
sc45	−0.0467	0.005

Partial correlation of breakfast with

Variable	Corr.	Sig.
sc12	0.0672	0.000
sc45	−0.0247	0.136

Partial correlation of sleepgd with

Variable	Corr.	Sig.
sc12	0.0423	0.011
sc45	0.0018	0.914

Partial correlation of alqprud with

Variable	Corr.	Sig.
sc12	−0.0425	0.010
sc45	−0.0312	0.059

Partial correlation of nobese with

Variable	Corr.	Sig.
sc12	0.0680	0.000
sc45	−0.0059	0.722

Partial correlation of exercise with

Variable	Corr.	Sig.
sc12	0.0655	0.000
sc45	−0.0659	0.000

The variable nsmoker is correlated with the occupational social class variables; the correlation is positive for sc12 and negative for sc45. Smoking is usually found to be more prevalent among the poorest individuals. The same pattern is observed in the correlation between breakfast, nobese and exercise, although the correlation coefficient for sc45 is statistically significant only for exercise. The variable sleepgd is positively correlated with sc12 and alqprud is negatively correlated with both the top and the bottom social class.

Correlation coefficients by education level are computed as well, holding lhqO constant:

```
• foreach x of global lifestyles!
                pcorr 'x' lhqdg lhqhndA lhqoth lhqnone
                }
```

Partial correlation of nsmoker with

Variable	Corr.	Sig.
lhqdg	0.0441	0.008
lhqhndA	0.0137	0.407
lhqoth	−0.0452	0.006
lhqnone	−0.0468	0.005

Partial correlation of breakfast with

Variable	Corr.	Sig.
lhqdg	0.0369	0.026
lhqhndA	−0.0004	0.980
lhqoth	−0.0274	0.098
lhqnone	−0.0422	0.011

Partial correlation of sleepgd with

Variable	Corr.	Sig.
lhqdg	−0.0012	0.944
lhqhndA	−0.0098	0.554
lhqoth	−0.0040	0.808
lhqnone	−0.0383	0.021

Partial correlation of alqprud with

Variable	Corr.	Sig.
lhqdg	−0.0027	0.871
lhqhndA	−0.0076	0.647
lhqoth	−0.0183	0.269
lhqnone	0.0156	0.345

Partial correlation of nobese with

Variable	Corr.	Sig.
lhqdg	0.0291	0.078
lhqhndA	0.0002	0.992
lhqoth	0.0136	0.410
lhqnone	−0.0230	0.164

Partial correlation of exercise with

Variable	Corr.	Sig.
lhqdg	−0.0070	0.673
lhqhndA	−0.0098	0.553
lhqoth	−0.0281	0.090
lhqnone	−0.1113	0.000

Negative correlations are found between nsmoker, breakfast, sleepgd as well as exercise and the less well-educated individuals.

In order to see how lifestyles are distributed across socioeconomic groups, we divide the sample into groups according to the number of 'healthy' behaviours adopted:

• gen ls1=nsmoker
  • gen ls2=breakfast
  • gen ls3=sleepgd
  • gen ls4=alqprud
  • gen ls5=nobese
  • gen ls6=exercise
  • summ ls1−ls6
  • egen sumls=rsum (ls1−ls6)
  • global zero "sumls==0"

• global one "sumls==1"
  • global two "sumls==2"
  • global three "sumls=3"
  • global four "sumls==4"
  • global five "sumls==5"
  • global six "sumls==6"

We compare the sample means of socioeconomic and demographic variables between the three sub-samples that are defined according to the number of healthy behaviours: 0–2, 3–5 and 6:

• global subvars “death sah sc12 sc3 sc45 lhqdg lhqhndA lhqO lhqnone lhqoth full part unemp sick retd keephse wkshft1 male age”

• sum $subvars if $zero | $one | $two
  • sum $subvars if $three | $four | $five
  • sum $subvars if $six

These are collected in the following table:

	Full sample	0/1/2	3/4/5	6
death	0.359	0.401	0.374	0.191
sah	0.703	0.605	0.696	0.863
sc12	0.316	0.215	0.306	0.500
sc3	0.467	0.497	0.473	0.380
sc45	0.218	0.290	0.220	0.123
lhqdg	0.125	0.078	0.118	0.237
lhqhnda	0.125	0.094	0.124	0.165
lhqo	0.094	0.059	0.093	0.143
lhqnone	0.608	0.716	0.619	0.402
lhqoth	0.047	0.054	0.046	0.054
full	0.364	0.462	0.343	0.436
part	0.132	0.116	0.124	0.217
unemp	0.030	0.054	0.030	0.009
sick	0.033	0.054	0.033	0.011
retd	0.339	0.191	0.372	0.219
keephse	0.102	0.124	0.098	0.108
wkshft1	0.057	0.102	0.055	0.031
male	0.455	0.505	0.451	0.442
age	57.468	53.889	58.412	53.380
N	3655	372	2932	351

The number of deaths decreases moving from the group with the fewest healthy behaviours to the group with the healthiest lifestyle. The more healthy behaviours there are, the bigger the proportion of persons belonging to the higher occupational social classes. The number of individuals in the bottom classes decreases moving from the most unhealthy lifestyles to the healthiest. The general result is that lifestyles are not randomly distributed but cluster together in certain groups of the population, suggesting that the relationship between lifestyle and socioeconomic environment must be taken into account. However, this does not say whether, and to what extent, health is affected by the propensities to undertake behaviours. Further analysis is needed.

Health and lifestyle

Using pcorr we also look at the partial correlation between lifestyles and health measures.

• pcorr sah $lifestyles
• pcorr death $lifestyles

Partial correlation of sah with

Variable	Corr.	Sig.
nsmoker	0.1057	0.000
breakfast	0.0127	0.443
sleepgd	0.0752	0.000
alqprud	−0.0255	0.124
nobese	0.0667	0.000
exercise	0.1239	0.000

Partial correlation of death with

Variable	Corr.	Sig.
nsmoker	−0.0353	0.033
breakfast	0.0491	0.003
sleepgd	−0.0813	0.000
alqprud	0.0037	0.821
nobese	−0.0050	0.761
exercise	−0.1737	0.000

sah is positively correlated with all the lifestyle indicators, with the exception of breakfast and alqprud. Also mortality is positively correlated with all the lifestyle indicators except alqprud and nobese. The variable breakfast is positively correlated with death: this result will be shown to hold also in the econometric model. A Pearson Chi-squared test is used to further investigate the relationship between health, mortality and lifestyle, represented in two-way tables:

• foreach x of global lifestyles!
tab ‘x’ sah, chi2
}
• foreach x of global lifestyles!
tab ‘x’ death, chi2
}

The typical Stata output and a summary of the results of the tests are reported below.

	death		
nsmoker	0	1	Total
0	676	422	1,098
1	1,666	891	2,557
Total	2,342	1,313	3,655

Pearson chi2 (1)=4.2961 Pr=0.038

	death		sah	
	χ-square	p-value	χ-square	p-value
nsmoker	4.296	0.038	44.437	0.000
breakfast	3.801	0.051	7.086	0.008
sleepgd	24.505	0.000	24.289	0.000
alqprud	1.059	0.303	1.381	0.240
nobese	0.848	0.357	18.101	0.000
exercise	114.663	0.000	68.837	0.000

The tests confirm the existence of a strong correlation between death and the lifestyles nsmoker, breakfast, sleepg and exercise. With respect to the analysis of partial correlations, the link between breakfast and sah is found to be highly statistically significant.

Mortality and socioeconomic status

The HALS provides information about the cause of death. Causes of death are coded using the ICD-9-CM diagnostic and procedure code system. Stata has a built-in command (icd9) to decode each specific cause of death. We generate 19 dummy variables for types of disease-specific mortality:

- icd9 clean ucause, dp
  - icd9 generate u1=ucause, range (001/139)
  - label var u1 “infectious and parastic dis”
  - icd9 generate u2=ucause, range (140/239)
  - label var u2 “neoplasms”
  - icd9 generate u3=ucause, range (240/279)
  - label var u3 “endocrine, nutritional and metabolic dis and immunity disorders”

- icd9 generate u4=ucause, range (280/289)
  - label var u4 “dis of the blood and blood-forming organs”
  - icd9 generate u5=ucause, range (290/319)
  - label var u5 “menatal disorder”
  - icd9 generate u6=ucause, range (320/389)
  - label var u6 “dis of the nervous sustem and sense organs”
  - icd9 generate u7=ucause, range (390/459)
  - label var u7 “dis of the circulatory system”

• icd9 generate u8=ucause, range (460/519)
• label var u8 “dis of the respiratory system”
• icd9 generate u9=ucause, range (520/579)
• label var u9 “dis of the digestive system”
• icd9 generate u10=ucause, range (580/629)
• label var u10 “dis of the genitourinary system”
• icd9 generate u11=ucause, range (630/679)
• label var u11 “complications of pregnancy, childbirth and the puerperium”
• icd9 generate u12=ucause, range (680/709)
• label var u12 “dis of the skin and subcutaneous tissue”
• icd9 generate u13=ucause, range (710/739)
• label var u13 “dis of the musculoskeletal system and connective tissue”
• icd9 generate u14=ucause, range (740/759)
• label var u14 “congenital anomalies”
• icd9 generate u15=ucause, range (760/779)
• label var u15 “certain conditions originating in the perinatal period”
• icd9 generate u16=ucause, range (780/799)
• label var u16 “symptoms, signs, and ill-defined conditions”
• icd9 generate u17=ucause, range (800/999)
• label var u17 “injury and poisoning”
• icd9 generate u18=ucause, range (E800/E999)
• label var u18 “supplementary classification of external causes of injury and poisoning”
• icd9 generate u19=ucause, range (V01/V83)
• label var u19 “supplementary classification of factors influencing health status and contact with health services”

• gen cd=0 /*cd=0 if missing cause*/
• replace cd=1 if u1==1
……
• replace cd=19 if u19==1

We explore the distribution of causes of death in the sample as a whole and split by social class (scgr equals 1 for sc12, 2 for sc3 and 3 for sc45):

• desc u1–u19
• tab cd if death==1

cd	Freq.	Percent	Cum.
0	27	2.06	2.06
1	9	0.69	2.74
2	367	27.95	30.69
3	17	1.29	31.99
4	2	0.15	32.14
5	20	1.52	33.66
6	15	1.14	34.81

7	597	45.47	80.27
8	162	12.34	92.61
9	43	3.27	95.89
10	15	1.14	97.03
13	4	0.30	97.33
16	13	0.99	98.32
17	6	0.46	98.78
18	16	1.22	100.00
Total	1,313	100.00	

• by scgr: tab cd if death=1

-> scgr=1

cd	Freq.	Percent	Cum.
0	7	2.26	2.26
1	4	1.29	3.55
2	89	28.71	32.26
3	5	1.61	33.87
4	2	0.65	34.52
5	10	3.23	37.74
6	3	0.97	38.71
7	126	40.65	79.35
8	35	11.29	90.65
9	9	2.90	93.55
10	6	1.94	95.48
16	5	1.61	97.10
17	3	0.97	98.06
18	6	1.94	100.00
Total	310	100.00	

-> scgr = 2

cd	Freq.	Perc ent	Cum.
0	15	2.26	2.26
1	3	0.45	2.71
2	191	28.81	31.52
3	5	0.75	32.28
5	6	0.90	33.18
6	10	1.51	34.69
7	319	48.11	82.81
8	72	10.86	93.67

9	17	2.56	96.23
10	8	1.21	97.44
13	4	0.60	98.04
16	5	0.75	98.79
17	3	0.45	99.25
18	5	0.75	100.00
Total	663	100.00	

-> scgr = 3

cd	Freq.	Percent	Cum.
0	5	1.47	1.47
1	2	0.59	2.06
2	87	25.59	27.65
3	7	2.06	29.71
5	4	1.18	30.88
6	2	0.59	31.47
7	152	44.71	76.18
8	55	16.18	92.35
9	17	5.00	97.35
10	1	0.29	97.65
16	3	0.88	98.53
18	5	1.47	100.00
Total	340	100.00	

The most frequent causes of death are diseases of the circulatory system (u7), neoplasms (u2) and diseases of the respiratory system (u8). Deaths in the three classes are mainly due to diseases of the respiratory system, with a maximum of 48% of deaths due to this cause among those in skilled occupations (sc3). The incidence of respiratory diseases is higher for semi and non-skilled occupations (sc45).

A crude way to see if mortality varies with the characteristics of the population is to look at the simple death rate. Stata has a built-in command called proportion, which produces estimates of the proportion of deaths by any covariate. We show how to calculate this directly:

• global varlist “sc12 sc3 sc45 lhqdg lhqhndA lhqO lhqnone male sah”
  • foreach x of global varlist{
    qui count if ‘x’==1&death==1
    scalar d ‘x’=r(N)
    disp “‘x’=1 and death=1:d ‘x’=” d ‘x’
    qui count if ‘x’==1
    scalar n ‘x’=r(N)
    disp “‘x’=1: n ‘x’=” n ‘x’
    scalar drate ‘x’=(d ‘x’/n ‘x’)*100

```
disp "death rate: d 'x'/n 'x'=" drate 'x'
disp " "
}
```

	Death rate
sc12	26.89
sc3	38.86
sc45	42.71
hqdg	24.73
lhqhnda	22.59
lhqo	22.32
lhqnone	43.10
male	42.91
sah	31.09

The death rate increases from the highest social class to the lowest. Such a clear gradient is not found across education levels. The mortality rate is higher for men and for individuals in fair or poor health status.

5.4 ESTIMATION STRATEGY AND RESULTS

This section illustrates our estimation strategy and describes the econometric approach adopted to obtain estimates of the effect of the lifestyle variables on health. We estimate a model that allows us to control for unobservable individual heterogeneity, which is a common methodological problem in the empirical estimation of the health production function.

The econometric model

The model described in (5.1) on p. 82 is a recursive triangular system of equations for lifestyles, morbidity and mortality. We assume that the random components of the lifestyle equations are correlated with the random components of the mortality and health equations. This means that potentially there are factors, unobservable to the researchers, that influence individual health-related behaviours as well as health status and the risk of mortality. Hence, the issue is to take into account this unobservable individual-specific heterogeneity in the estimation procedure in order to recover consistent estimates of the coefficients. Potential endogeneity of self-assessed health and the lifestyle variables in the recursive model is reflected in the correlation between the error terms and the exogenous covariates as well as in the correlation between disturbances of all the equations of the model. If endogeneity is proven to be a problem, then coefficient estimates from a univariate probit model for mortality will be inconsistent.

In this analysis, a multivariate probit model is used for estimation because it not only allows for dependence and deals appropriately with unobservable heterogeneity and potential endogeneity, but gives mortality a structural representation in the model. Other

empirical work has used various single-equation methods, capturing endogeneity with the method of two-stage least-squares and the generalized method of moments (see, e.g., Auster *et al.* 1969; Rosenzweig and Schultz 1983; Grossman and Joyce 1990; Mullahy and Portney 1990; Mullahy and Sindelar 1996).

We estimate a triangular recursive system, which consists of structural equations for the health production functions and six reduced-form equations for lifestyles. The dependent variables in the recursive model are binary variables: y_{id}, y_{ih} and y_{ic} denote death, sah and the set of lifestyles (nsmoker breakfast sleepgd alqprud nobese exercise) respectively. The latent variables underlying each observed variable define the following system of equations:

$$\begin{aligned} y^*_{id} &= \boldsymbol{\delta}'_d \boldsymbol{y}_{ic} + \boldsymbol{\vartheta}'_d y_{ih} + \boldsymbol{\alpha}'_d \boldsymbol{w}_i + \varepsilon_{id} \\ y^*_{ih} &= \boldsymbol{\delta}'_h \boldsymbol{y}_{ic} + \boldsymbol{\alpha}'_h \boldsymbol{w}_i + \boldsymbol{\beta}'_h \boldsymbol{z}_i + \varepsilon_{ih} \\ \boldsymbol{y}^*_{ic} &= \boldsymbol{\alpha}'_c \boldsymbol{w}_i + \boldsymbol{\beta}'_c \boldsymbol{z}_i + \boldsymbol{\gamma}_c \boldsymbol{v}_i + \varepsilon_{ic} \end{aligned} \tag{5.2}$$

such that

$$\begin{aligned} y_{id} &= 1(y^*_{id} \geq 0) \\ y_{ih} &= 1(y^*_{ih} \geq 0) \\ \boldsymbol{y}_{ic} &= 1(\boldsymbol{y}^*_{ic} \geq 0) \end{aligned}$$

where $y_{ic}=\{y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6}\}$ is a vector of six lifestyles, and w_i, z_i and v_i are individual-specific exogenous vectors that explain, respectively, mortality, health and lifestyle; health and lifestyle; and lifestyle only. These variables are chosen according to an approach to identification that will be illustrated later.

The error terms of the latent equations have a multivariate normal distribution, giving a multivariate probit model. The log-likelihood depends on a multivariate standard normal distribution, ϕ_M. Full information maximum likelihood (FIML) estimation cannot be performed directly as the integrals in the likelihood function have no closed form. Therefore the model is estimated using the command mvprobit written for Stata by Cappellari and Jenkins, which uses the GHK (Geweke-Hajivassilou-Keane) simulator for probabilities and a maximum simulated likelihood (MSL) procedure. The algorithm implemented in mvprobit is described in Cappellari and Jenkins (2003).

The GHK simulator exploits the Choleski decomposition of the covariance matrix, so that the joint probability originally based on unobservables can be written as the product of univariate conditional probabilities, where the errors in the M equations are substituted by disturbances that are independent of each other by construction. A maximum likelihood procedure using the GHK simulator at each iteration is numerically intensive and simulation bias may arise. For further details about MSL see Contoyannis, Jones and Leon-Gonzalez (2004) and Train (2003).

Estimation

Systems of binary dependent variables with endogenous binary regressors typically require exclusion restrictions for robust identification of the parameters (see Maddala (1983) for a more detailed insight). This would-imply finding a set of variables that are instrumental to identify the effect of y_{ih} and y_{il} in the mortality, and health and mortality equations, respectively, as shown in (5.2). However, given the assumption of joint normality, the model is identified by functional form, which does not require any exclusion restrictions, such that the regressors for all the M equations are identical (see Wilde 2000). We compare results for models with and without exclusion restrictions.

We choose the best set of exclusion restrictions looking at the statistical fit of different specifications. In particular, both in the univariate mortality equation probit and in the multivariate probit model, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) support the exclusion of wales north nwest yorks wmids emids anglia swest london scot widow divorce seprd single housown hou smother mothsmo fathsmo bothsmo alpa alma from the mortality equations. The same test is used to compare this model with a model estimated without exclusion restrictions.

We first define the right-hand sides of our equations, such that deq heq and leq are the regressors for mortality, health and lifestyles, respectively, when exclusion restrictions are set; xvars are the regressors for each equation when no restriction is required and deqex is the right-hand side for the exogenous model, where only exogenous regressors are included. We include a squared term for age, which allows the probability of death to be a smooth and flexible function of age.

global deq “sah nsmoker breakfast sleepgd alqprud nobese exercise sc12 sc45 lhqdg lhqhndA lhqnone lhqoth part unemp sick retd keephse wkshft1 rural suburb ethwheur height male age age2”

• global heq “nsmoker breakfast sleepgd alqprud nobese exercise sc12 sc45 lhqdg lhqhndA lhqnone lhqoth widow divorce seprd single part unemp sick retd keephse wkshf t1 wales north nwest yorks wmids emids anglia swest london scot rural suburb ethwheur housown hou height male age age2”

• global lseq “sc12 sc45 lhqdg lhqhndA lhqnone lhqoth widow divorce seprd single part unemp sick retd keephse wkshf t1 wales north nwest yorks wmids emids anglia swest london scot rural suburb ethwheur housown hou height male age age2 smother mothsmo fathsmo bothsmo alpa alma”

• global deqex “sc12 sc45 lhqdg lhqhndA lhqnone lhqoth part unemp sick retd keephse wkshf t1 rural suburb ethwheur height male age age2”

• global xvars “sah nsmoker breakfast sleepgd alqprud nobese exercise sc12 sc45 lhqdg lhqhndA lhqnone lhqoth widow divorce seprd single part unemp sick retd keephse wkshf t1 wales north nwest yorks wmids emids anglia swest london scot rural suburb ethwheur housown hou height male age age2 smother mothsmo fathsmo bothsmo alpa alma”

We estimate univariate probit models for mortality using the command probit, and compare the two identification approaches using the post-estimation command fits tat. The command fits tat computes fit statistics for single-equation regression models and

can be downloaded by typing findit fitstat in Stata. We also test mis-specification of the models by means of the RESET test:

- probit death $deq/*, nolog*/
  - fitstat, saving (m1)
  - /*RESET test*/
  - predict yhat, xb
  - gen yhat2=yhat^2
  - qui probit death $deqex yhat2/*, nolog*/
  - test yhat2=0
  - drop yhat yhat2

- probit death $xvars/*, nolog*/
  - fitstat, using (m1)
  - /*RESET test*/
  - predict yhat, xb
  - gen yhat2=yhat^2
  - qui probit death $xvars yhat2/*, nolog*/
  - test yhat2=0
  - drop yhat yhat2

Tables 5.1 and 5.2 report the results of the estimation:

Table 5.1 Probit model for mortality—with exclusion restrictions

Probit regression				Number of obs=	3655
				LR chi2 (26)=	1622.36
				Prob>chi2=	0.0000
Log likelihood =−1575.4469				Pseudo R2	=0.3399
death	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
sah	−.3014813	.0570352	−5.29	0.000	−.4132682 −.1896944
nsmoker	−.3539883	.0587143	−6.03	0.000	−.4690662 −.2389104
breakfast	−.1545135	.059715	−2.59	0.010	−.2715527 −.0374743
sleepgd	−.0794598	.0519173	−1.53	0.126	−.1812159 .0222963
alqprud	−.1164528	.0809028	−1.44	0.150	−.2750194 .0421138
nobese	−.1767824	.0715364	−2.47	0.013	−.3169912 −.0365736
exercise	−.090148	.0584445	−1.54	0.123	−.2046971 .0244012
sc12	−.1290353	.0668622	−1.93	0.054	−.2600828 .0020121
sc45	−.016304	.0648742	−0.25	0.802	−.1434551 .1108471
lhqdg	.0347702	.1225396	0.28	0.777	.205403 .2749435
lhqhndA	−.0833388	.1196754	−0.70	0.486	−.3178983 .1512208
lhqnone	.090173	.0986109	0.91	0.360	−.1031008 .2834468

lhqoth	−.0821548	.1489644	−0.55	0.581	−.3741196	.2098101
part	.1461972	.1014417	1.44	0.150	−.0526248	.3450192
unemp	.2904067	.1415827	2.05	0.040	0129097	.5679036
sick	.6038919	.1397277	4.32	0.000	.3300306	.8777531
retd	.0654	.0925226	0.71	0.480	−.1159409	.2467409
keephse	.24926	.1075017	2.32	0.020	0385605	.4599596
wkshft	−.2638565	.129371	−2.04	0.041	−.517419	−.010294
rural	−.1600735	.0735798	−2.18	0.030	−.3042872	−.0158597
suburb	−.0721698	.0589076	−1.23	0.221	−.1876265	.0432869
ethwheur	.3997708	.202348	1.98	0.048	.003176	.7963656
height	.010011	.0096029	1.04	0.297	−.0088104	.0288324
male	.4333059	.0789582	5.49	0.000	.2785507	.5880611
age	.0358631	.0261894	1.37	0.171	−.0154673	.0871934
age2	.0395356	.0217351	1.82	0.069	−.0030645	.0821356
_cons	−4.291507	1.016465	−4.22	0.000	−6.283741	−2.299273

Measures of Fit for probit of death

Log-Lik Intercept Only:	−2386. 628	Log-Lik Full Model:	−1575.447
D(3628):	3150.894	LR(26):	1622.362
		Prob>LR:	0.000
McFadden's R2:	0.340	McFadden's Adj R2:	0.329
Maximum Likelihood R2:	1.000	Cragg & Uhler's R2:	1.000
McKelvey and Zavoina's R2:	0.537	Efron's R2:	0.400
Variance of y*:	2.159	Variance of error:	1.000
Count R2:	0.804	Adj Count R2:	0.455
AIC:	0.877	AIC*n:	3204.894
BIG:	−26612.679	BIG':	−1409.062

(Indices saved in matrix fs_m1)
The RESET test result is:
(1) yhat2=0
chi2 (1)=1.23
Prob > chi2=0.2671

Table 5.2 Probit model for mortality—without exclusion restrictions

Probit regression		Number of obs=	3655
		LR chi2 (48)=	1643.22
		Prob>chi2=	0.0000

Log likelihood =−1565.0185				Pseudo R2=		0.3443
death	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sah	−.2981329	.0575411	−5.18	0.000	−.4109113	−.1853544
nsmoker	−.3210206	.0608401	−5.28	0.000	−.440265	−.2017761
breakfast	−.1669265	.0605715	−2.76	0.006	−.2856445	−.0482084
sleepgd	−.0747964	.0524052	−1.43	0.154	−.1775087	.0279158
alqprud	−.0879154	.0823417	−1.07	0.286	−.2493022	.0734715
nobese	−.1655968	.0722707	−2.29	0.022	−.3072448	−.0239488
exercise	−.0906202	.0590312	−1.54	0.125	−.2063193	.0250788
sc12	−.1163491	.0673427	−1.73	0.084	−.2483385	.0156402
sc45	−.0238404	.065503	−0.36	0.716	−.1522239	.1045432
lhqdg	.0554918	.1243229	0.45	0.655	−.1881765	.2991602
lhqhndA	−.0619253	.1211734	−0.51	0.609	−.2994207	.1755702
lhqnone	.0985861	.1000026	0.99	0.324	−.0974155	.2945877
lhqoth	−.0762203	.1505662	−0.51	0.613	−.3713245	.218884
widow	.0913336	.0886763	1.03	0.303	−.0824688	.265136
divorce	.0307983	.144219	0.21	0.831	−.2518657	.3134624
seprd	.139194	.2153943	0.65	0.518	−.282971	.561359
single	.1220696	.1183595	1.03	0.302	−.1099107	.3540499
part	.1568007	.1027052	1.53	0.127	−.0444977	.3580992
unemp	.2556767	.143247	1.78	0.074	−.0250824	.5364357
sick	.5688889	.1411486	4.03	0.000	.2922427	.8455352
retd	.0662907	.0935584	0.71	0.479	−.1170804	.2496619
keephse	.2639528	.1093166	2.41	0.016	.0496963	.4782093
wkshft1	−.2654164	.1304876	−2.03	0.042	−.5211673	−.0096655
wales	.1381786	.1276515	1.08	0.279	−.1120137	.3883709
north	.3061477	.1180372	2.59	0.009	.074799	.5374965
nwest	.2210419	.0955407	2.31	0.021	.0337856	.4082982
yorks	.0979064	.1099286	0.89	0.373	−.1175498	.3133625
wmids	.2569787	.1094335	2.35	0.019	.0424929	.4714644
emids	.1302095	.1122042	1.16	0.246	−.0897067	.3501256
anglia	−.0437136	.1447139	−0.30	0.763	−.3273477	.2399204
swest	.2287135	.1093917	2.09	0.037	.0143097	.4431172
london	.1931736	.1058563	1.82	0.068	−.014301	.4006482
scot	.2605411	.1042075	2.50	0.012	.0562983	.464784
rural	−.109596	.0775725	−1.41	0.158	−.2616353	.0424433
suburb	−.0386666	.0606252	−0.64	0.524	−.1574897	.0801566
ethwheur	.3409348	.2068292	1.65	0.099	−.064443	.7463125
housown	.0281038	.1542335	0.18	0.855	−.2741883	.330396
hou	.005987	.0301523	0.20	0.843	−.0531104	.0650845

death	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
height	.0141288	.0097319	1.45	0.147	−.0049453	.0332029
male	.4349046	.0810094	5.37	0.000	.276129	.5936801
age	.0398603	.0279627	1.43	0.154	−.0149455	.0946662
age2	.0371239	.0228867	1.62	0.105	−.0077333	.081981
smother	.0700997	.0615552	1.14	0.255	−.0505462	.1907456
mothsmo	−.0161021	.1716008	−0.09	0.925	−.3524335	.3202293
fathsmo	.0930378	.0814721	1.14	0.253	−.0666446	.2527202
bothsmo	.1044291	.0972164	1.07	0.283	−.0861115	.2949697
alpa	.0164571	.0230832	0.71	0.476	−.028785	.0616993
alma	−.001592	.0285795	−0.06	0.956	−.0576067	.0544227
_cons	−5.102805	1.118332	−4.56	0.000	−7.294695	−2.910914

Measures of Fit for probit of death

Model:	Current probit	Saved probit	Difference
N:	3655	3655	0
Log-Lik Intercept Only:	−2386.628	−2386.628	0.000
Log-Lik Full Model:	−1565.019	−1575.447	10.428
D:	3130.037 (3606)	3150.894 (3628)	−20.857 (−22)
LR:	1643.219 (48)	1622.362 (26)	20.857 (22)
Prob>LR:	0.000	0.000	0.000
McFadden's R2:	0.344	0.340	0.004
McFadden's Adj R2:	0.324	0.329	−0.005
Maximum Likelihood R2:	1.000	1.000	0.000
Cragg & Uhler's R2:	1.000	1.000	0.000
McKelvey and Zavoina's R2:	0.544	0.537	0.007
Efron's R2:	0.405	0.400	0.005
Variance of y*:	2.192	2.159	0.034
Variance of error:	1.000	1.000	0.000
Count R2:	0.808	0.804	0.004
Adj Count R2:	0.465	0.455	0.010
AIC:	0.883	0.877	0.006
AIC*n:	3228.037	3204.894	23.143
BIC:	−26453.051	−26612.679	159.628
BIG':	−1249.434	−1409.062	159.628

Difference of 159.628 in BIC' provides very strong support for saved model.

The RESET test for this specification of the model results in:

(1) yhat2=0

chi2 (1)=1.65

Prob>chi2=0.1986

In fitstat the AIC is calculated as $AIC=(-2\log L+2q)/N$, which is the per-observation contribution to the penalized likelihood. The smaller the AIC and BIC are, the better the fit of the model. Here they both favour the model with exclusion restrictions. The RESET test suggests that the mortality equation is not mis-specified in both cases.

Likewise, we try to estimate and compare two different specifications of the eight-dimensional multivariate probit model. However, only the model with exclusion restrictions converges to a global maximum.

We first install the module by typing:

• ssc install mvprobit

Then run the following model, specifying the option draws (#) for the number of random draws in the simulation procedure (higher (#) increases accuracy but is more time-consuming) and setting memory and mat size accordingly (a good rule of thumb is to set memory at least equal to the value (#)*M, and increase matsize for models with many covariates):

• set memory 400

• set matsize 10000

• mvprobit (death=$deq) (sah=$heq) (nsmoker=$leq) (breakfast=$leq) (sleepgd=$leq) (alqprud=$leq) (nobese=$leq) (exercise=$leq), dr(50)

Table 5.3 reports a part of the Stata output.

Table 5.3 Multivariate probit—8 equations

Multivariate probit (SML, # draws=50)				Number of obs=		3655
				Wald chi2 (313)=		3813.39
Log likelihood=−14535.445				Prob>chi2=		0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
death						
sah	−.5821151	.2964812	−1.96	0.050	−1.163208	−.0010227
nsmoker	−.8753122	.1603157	−5.46	0.000	−1.189525	−.5610991
breakfast	.4694373	.1980918	2.37	0.018	.0811845	.8576901
sleepgd	−.6965091	.234478	−2.97	0.003	−1.156077	−.2369408
alqprud	−.1668721	.3233303	−0.52	0.606	−.8005878	.4668435
nobese	.1772641	.2380997	0.74	0.457	−.2894027	.6439308
exercise	.0782	.2410795	0.32	0.746	−.3943072	.5507072
sc12	−.0803477	.0676633	−1.19	0.235	−.2129653	.0522699

sc45	−.0315815	.0636424	−0.50	0.620	−.1563183	.0931553
lhqdg	.0392397	.1136549	0.35	0.730	−.1835197	.2619992
lhqhndA	−.0493236	.1109303	−0.44	0.657	−.266743	.1680958
lhqnone	.0774288	.0986621	0.78	0.433	−.1159453	.2708029
lhqoth	−.0786728	.1396297	−0.56	0.573	−.3523419	.1949963
part	.151256	.0989818	1.53	0.126	−.0427448	.3452567
unemp	.26476	.1374437	1.93	0.054	−.0046247	.5341446
sick	.2772725	.2238468	1.24	0.215	−.1614592	.7160042
retd	−.0193505	.0941974	−0.21	0.837	−.203974	.1652731
keephse	.2578876	.1056766	2.44	0.015	.0507652	.4650099
wkshft1	−.2722732	.1222796	−2.23	0.026	−.5119369	−.0326095
rural	−.1054508	.0726564	−1.45	0.147	−.2478548	.0369532
suburb	−.1005662	.0572258	−1.76	0.079	−.2127267	.0115943
ethwheur	.3030995	.2033053	1.49	0.136	−.0953715	.7015705
height	.0090763	.0090375	1.00	0.315	−.008637	.0267895
male	.3360145	.0870288	3.86	0.000	.1654412	.5065878
age	.0065637	.0259159	0.25	0.800	−.0442306	.057358
age2	.0553238	.0213923	2.59	0.010	.0133957	.0972518
_cons	−2.679849	1.066136	−2.51	0.012	−4.769437	−.5902616
sah						
nsmoker	.0426545	.1823199	0.23	0.815	−.314686	.399995
breakfast	.3042656	.2680987	1.13	0.256	−.2211981	.8297294
sleepgd	.0259409	.2924361	0.09	0.929	−.5472234	.5991052
alqprud	−.0358838	.3148629	−0.11	0.909	−.6530038	.5812362
nobese	.7823743	.2632659	2.97	0.003	.2663825	1.298366
exercise	.2426147	.3497862	0.69	0.488	−.4429537	.9281832
sc12	.1700142	.0637461	2.67	0.008	.0450742	.2949542
sc45	−.1362953	.0599449	−2.27	0.023	−.2537852	−.0188055
lhqdg	.114918	.1080193	1.06	0.287	−.0967959	.3266319
lhqhndA	.0668313	.1025091	0.65	0.514	−.1340829	.2677455
lhqnone	−.1332119	.0957892	−1.39	0.164	−.3209552	.0545314
lhqoth	−.0830815	.1321294	−0.63	0.529	−.3420504	.1758873
widow	−.1407317	.0810156	−1.74	0.082	−.2995194	.018056
divorce	−.1108505	.126455	−0.88	0.381	−.3586978	.1369968
seprd	−.0069978	.1846403	−0.04	0.970	−.3688863	.3548906
single	.0132986	.1139187	0.12	0.907	−.2099779	.2365752
part	−.0621853	.0937114	−0.66	0.507	−.2458562	.1214856
unemp	−.1632309	.1379538	−1.18	0.237	−.4336154	.1071536
sick	−1.532382	.1855145	−8.26	0.000	−1.895984	−1.16878
retd	−.2278248	.0915595	−2.49	0.013	−.4072781	−.0483716

keephse	−.2325369	.0967617	−2.40	0.016	−.4221864	−.0428875
wkshft1	.0264542	.1121714	0.24	0.814	−.1933977	.246306
wales	−.2290903	.1110686	−2.06	0.039	−.4467808	−.0113997
north	−.0158852	.1133259	−0.14	0.889	−.2379999	.2062296
nwest	−.1974732	.0883151	−2.24	0.025	−.3705677	−.0243787
yorks	−.0915665	.0951662	−0.96	0.336	−.2780888	.0949559
wmids	−.0810523	.10273	−0.79	0.430	−.2823994	.1202947
emids	.0103476	.1010353	0.10	0.918	−.1876779	.2083731
anglia	−.1210304	.1257659	−0.96	0.336	−.3675271	.1254662
swest	−.1444995	.0950496	−1.52	0.128	−.3307934	.0417944
london	−.0728538	.0955833	−0.76	0.446	−.2601936	.1144861
scot	−.1989489	.0978494	−2.03	0.042	−.3907303	−.0071676
rural	.092393	.0726131	1.27	0.203	−.0499262	.2347121
suburb	.0205645	.0602075	0.34	0.733	−.0974401	.1385691
ethwheur	.4095968	.1806043	2.27	0.023	.0556189	.7635748
housown	−.1876902	.1403905	−1.34	0.181	−.4628506	.0874701
hou	.010681	.0292779	0.36	0.715	−.0467026	.0680645
height	−.004792	.0087212	−0.55	0.583	−.0218852	.0123012
male	−.0577633	.0868588	−0.67	0.506	−.2280035	.1124768
age	−.0427561	.0237577	−1.80	0.072	−.0893204	.0038082
age2	.0362047	.0181426	2.00	0.046	.0006459	.0717634
_cons	1.187633	1.087138	1.09	0.275	−.9431188	3.318385
nsmoker						
sc12	.1428484	.0598786	2.39	0.017	.0254885	.2602082
sc45	−.085165	.0587121	−1.45	0.147	−.2002386	.0299086
lhqdg	.1087093	.1073014	1.01	0.311	−.1015975	.3190161
lhqhndA	−.0504252	.1013647	−0.50	0.619	−.2490963	.1482458
lhqnone	−.210612	.0840795	−2.50	0.012	−.3754047	−.0458192
lhqoth	−.3465579	.1281228	−2.70	0.007	−.5976741	−.0954418
widow	−.2303274	.0839917	−2.74	0.006	−.3949481	−.0657068
divorce	−.3742869	.1202567	−3.11	0.002	−.6099857	−.138588
seprd	−.2276467	.1758575	−1.29	0.195	−.572321	.1170276
single	−.160793	.1101347	−1.46	0.144	−.376653	.0550669
part	−.1069453	.0823294	−1.30	0.194	−.2683079	.0544174
unemp	−.3854984	.1317899	−2.93	0.003	−.643802	−.1271949
sick	−.2651143	.1316715	−2.01	0.044	−.5231858	−.0070428
retd	−.2771589	.0894183	−3.10	0.002	−.4524156	−.1019022
keephse	−.1305489	.0906925	−1.44	0.150	−.3083029	.0472051
wkshft1	−.1732417	.0990383	−1.75	0.080	−.3673532	.0208697
wales	−.2017046	.1096867	−1.84	0.066	−.4166866	.0132774

north	−.3346367	.1035359	−3.23	0.001	−.5375633	−.13171
nwest	−.3234131	.0834084	−3.88	0.000	−.4868906	−.1599355
yorks	−.246909	.0950004	−2.60	0.009	−.4331063	−.0607116
wmids	−.2461482	.0974849	−2.52	0.012	−.437215	−.0550814
emids	−.0823298	.1020729	−0.81	0.420	−.2823891	.1177295
anglia	−.0783644	.128174	−0.61	0.541	−.3295809	.172852
swest	−.0851	.0979412	−0.87	0.385	−.2770613	.1068613
london	−.1706743	.0956462	−1.78	0.074	−.3581374	.0167889
scot	−.373766	.0920466	−4.06	0.000	−.5541741	−.1933578
rural	.0924518	.0691197	1.34	0.181	−.0430204	.227924
suburb	.0199923	.0544614	0.37	0.714	−.08675	.1267346
ethwheur	.0713028	.1636629	0.44	0.663	−.2494707	.3920762
housown	−.0897821	.1358694	−0.66	0.509	−.3560812	.176517
hou	.0508187	.0248557	2.04	0.041	.0021024	.099535
height	.0093824	.008836	1.06	0.288	−.0079358	.0267006
male	−.2464529	.0723612	−3.41	0.001	−.3882782	−.1046275
age	−.0494993	.0225196	−2.20	0.028	−.0936369	−.0053617
age2	.0624777	.0190546	3.28	0.001	.0251314	.099824
smother	−.7094384	.0510863	−13.89	0.000	−.8095657	−.6093112
mothsmo	−.4352468	.1458825	−2.98	0.003	−.7211712	−.1493223
fathsmo	−.1887278	.0770558	−2.45	0.014	−.3397544	−.0377012
bothsmo	−.2755882	.0874484	−3.15	0.002	−.446984	−.1041924
alpa	−.0456466	.0206769	−2.21	0.027	−.0861727	−.0051206
alma	−.0466048	.0251863	−1.85	0.064	−.0959691	.0027595
_cons	1.762126	.9114654	1.93	0.053	−.0243134	3.548565
breakfast						
sc12	.0520441	.0588782	0.88	0.377	−.063355	.1674433
sc45	−.0457869	.0585542	−0.78	0.434	−.160551	.0689773
lhqdg	.1830792	.1055484	1.73	0.083	−.0237919	.3899503
lhqhndA	−.0528413	.1002552	−0.53	0.598	−.2493379	.1436554
lhqnone	−.2732616	.08373	−3.26	0.001	−.4373694	−.1091538
lhqoth	−.2218262	.1285937	−1.73	0.085	−.4738651	.0302127
widow	−.1862163	.0844298	−2.21	0.027	−.3516955	−.020737
divorce	−.2686231	.118006	−2.28	0.023	−.4999105	−.0373357
seprd	−.3355242	.172983	−1.94	0.052	−.6745648	.0035163
single	−.0926959	.1098038	−0.84	0.399	−.3079073	.1225155
part	.1596601	.0820596	1.95	0.052	−.0011737	.320494
unemp	−.2986676	.1287864	−2.32	0.020	−.5510844	−.0462509
sick	−.2098891	.1261561	−1.66	0.096	−.4571504	.0373723
retd	.2353614	.0890724	2.64	0.008	.0607828	.40994

keephse	−.0486515	.0883067	−0.55	0.582	−.2217293	.1244264
wkshft1	−.1507742	.0962189	−1.57	0.117	−.3393597	.0378112
wales	.1122796	.1074019	1.05	0.296	−.0982242	.3227835
north	.1694127	.1049648	1.61	0.107	−.0363145	.3751398
nwest	.2426892	.0847878	2.86	0.004	.0765081	.4088703
yorks	−.0066566	.092545	−0.07	0.943	−.1880415	.1747282
wmids	.0067598	.0937124	0.07	0.942	−.1769131	.1904326
emid	−.0611793	.0962121	−0.64	0.525	−.2497516	.127393
anglia	−.0220602	.1256606	−0.18	0.861	−.2683504	.2242301
swest	.0681359	.0944055	0.72	0.470	−.1168954	.2531672
london	.0702718	.0930653	0.76	0.450	−.1121328	.2526763
scot	.2156251	.0929975	2.32	0.020	.0333534	.3978969
rural	.1188946	.0677851	1.75	0.079	−.0139618	.251751
suburb	.1658145	.0540403	3.07	0.002	.0598974	.2717316
ethwheur	.4769197	.1519494	3.14	0.002	.1791043	.7747352
housown	−.143907	.1335708	−1.08	0.281	−.4057009	.1178869
hou	−.0284954	.0242925	−1.17	0.241	−.0761078	.0191169
height	.0158143	.0087835	1.80	0.072	−.0014011	.0330297
male	−.1157835	.0719798	−1.61	0.108	−.2568613	.0252944
age	.0240544	.0208576	1.15	0.249	−.0168258	.0649346
age2	−.0094758	.01741	−0.54	0.586	−.0435988	.0246471
smother	−.3420613	.0503755	−6.79	0.000	−.4407956	−.2433271
mothsmo	−.1743186	.1402961	−1.24	0.214	−.4492938	.1006567
fathsmo	.0731555	.0727794	1.01	0.315	−.0694896	.2158005
bothsmo	−.1269888	.0837008	−1.52	0.129	−.2910394	.0370619
alpa	−.0466668	.0204817	−2.28	0.023	−.0868102	−.0065235
alma	−.0224937	.0250663	−0.90	0.370	−.0716226	.0266353
_cons	−1.561126	.8737996	−1.79	0.074	−3.273741	.1514901
sleepgd						
sc12	.0325705	.0544401	0.60	0.550	−.0741302	.1392712
sc45	.0212167	.0552386	0.38	0.701	−.0870489	.1294823
lhqdg	.0378079	.095028	0.40	0.691	−.1484436	.2240594
lhqhndA	−.0474675	.0924274	−0.51	0.608	−.2286219	.1336869
lhqnone	−.0734149	.0776403	−0.95	0.344	−.2255872	.0787574
lhqoth	.0536007	.1213894	0.44	0.659	−.1843181	.2915195
widow	−.2003282	.07398	−2.71	0.007	−.3453263	−.05533
divorce	−.2508919	.1119326	−2.24	0.025	−.4702758	−.031508
seprd	−.3306915	.1639817	−2.02	0.044	−.6520896	−.0092933
single	−.1932351	.0996077	−1.94	0.052	−.3884625	.0019924
part	.316822	.0774297	4.09	0.000	.1650626	.4685815

unemp	.2301242	.130661	1.76	0.078	−.0259668	.4862151
sick	−.2051607	.1237005	−1.66	0.097	−.4476092	.0372879
retd	.1451823	.0805469	1.80	0.071	−.0126867	.3030514
keephse	.1952964	.0853192	2.29	0.022	.0280738	.3625191
wkshft1	−.2813357	.0942974	−2.98	0.003	−.4661552	−.0965162
wales	.0273016	.1006122	0.27	0.786	−.1698946	.2244978
north	−.142438	.0966138	−1.47	0.140	−.3317976	.0469216
nwest	.1162727	.0779613	1.49	0.136	−.0365287	.269074
yorks	.0045772	.0861907	0.05	0.958	−.1643536	.173508
wmids	−.072364	.0889452	−0.81	0.416	−.2466935	.1019655
emids	−.0862544	.0898216	−0.96	0.337	−.2623015	.0897927
anglia	−.0928191	.1148296	−0.81	0.419	−.317881	.1322427
swest	.0011307	.0869289	0.01	0.990	−.1692468	.1715082
london	−.0308947	.0851117	−0.36	0.717	−.1977105	.1359211
scot	−.1626501	.0842248	−1.93	0.053	−.3277276	.0024274
rural	.0576153	.063571	0.91	0.365	−.0669817	.1822122
suburb	−.0421899	.0505699	−0.83	0.404	−.1413051	.0569252
ethwheur	.35194	.15348	2.29	0.022	.0511248	.6527552
housown	−.5074553	.1273436	−3.98	0.000	−.7570441	−.2578665
hou	−.0210648	.0233251	−0.90	0.366	−.0667812	.0246516
height	.0029364	.0081041	0.36	0.717	−.0129474	.0188202
male	.0475561	.06553	0.73	0.468	−.0808803	.1759925
age	−.0247637	.018655	−1.33	0.184	−.0613268	.0117993
age2	.0106196	.0151697	0.70	0.484	−.0191125	.0403516
smother	−.0394695	.0481566	−0.82	0.412	−.1338547	.0549158
mothsmo	.2511501	.1369181	1.83	0.067	−.0172044	.5195046
fathsmo	.0868155	.0663324	1.31	0.191	−.0431936	.2168245
bothsmo	.0540347	.0783545	0.69	0.490	−.0995373	.2076068
alpa	−.0177333	.0187868	−0.94	0.345	−.0545548	.0190882
alma	.0059008	.0233876	0.25	0.801	−.039938	.0517397
_cons	1.225414	.8071682	1.52	0.129	−.3566067	2.807434
alqprud						
sc12	−.249019	.0756193	−3.29	0.001	−.3972301	−.1008078
sc45	−.1779298	.0769656	−2.31	0.021	−.3287796	−.0270801
lhqdg	.0147554	.1272541	0.12	0.908	−.2346581	.2641688
lhqhndA	−.1121106	.1232721	−0.91	0.363	−.3537195	.1294984
lhqnone	.0202595	.105356	0.19	0.848	−.1862345	.2267535
lhqoth	−.044109	.1561975	−0.28	0.778	−.3502504	.2620325
widow	−.051605	.1171296	−0.44	0.660	−.2811748	.1779648
divorce	.098765	.1588266	0.62	0.534	−.2125294	.4100594

seprd	−.0894571	.2131938	−0.42	0.675	−.5073092	.328395
single	−.060026	.1384448	−0.43	0.665	−.3313727	.2113207
part	.2182917	.1136419	1.92	0.055	−.0044423	.4410256
unemp	−.1961343	.1453778	−1.35	0.177	−.4810696	.088801
sick	.4108216	.1734473	2.37	0.018	.0708712	.7507721
retd	.1957875	.1127076	1.74	0.082	−.0251154	.4166904
keephse	.1736469	.1302613	1.33	0.183	−.0816604	.4289543
wkshft1	−.0833196	.1151653	−0.72	0.469	−.3090394	.1424002
wales	−.2064332	.1444564	−1.43	0.153	−.4895625	.076696
north	−.5218218	.1338972	−3.90	0.000	−.7842554	−.2593882
nwest	−.2337961	.111753	−2.09	0.036	−.4528278	−.0147643
yorks	−.3669487	.1209438	−3.03	0.002	−.6039942	−.1299031
wmids	−.1224667	.1297098	−0.94	0.345	−.3766931	.1317598
emids	−.2997976	.1261441	−2.38	0.017	−.5470356	−.0525597
anglia	−.1001202	.1697305	−0.59	0.555	−.4327858	.2325454
swest	−.2933091	.1231661	−2.38	0.017	−.5347101	−.051908
london	−.2213895	.1224835	−1.81	0.071	−.4614527	.0186738
scot	−.3066555	.1225817	−2.50	0.012	−.5469113	−.0663997
rural	.1703	.0884087	1.93	0.054	−.0029778	.3435778
suburb	.0874449	.0688495	1.27	0.204	−.0474976	.2223874
ethwheur	−.5832431	.2922673	−2.00	0.046	−1.156076	−.0104098
housown	.1841803	.1625342	1.13	0.257	−.1343808	.5027415
hou	.1132175	.0319776	3.54	0.000	.0505425	.1758924
height	−.001065	.0111641	−0.10	0.924	−.0229462	.0208162
male	−.6977769	.0931732	−7.49	0.000	−.8803931	−.5151607
age	.0321039	.0271827	1.18	0.238	−.0211731	.085381
age2	−.0127048	.0231069	−0.55	0.582	−.0579935	.0325839
smother	−.2690448	.0653551	−4.12	0.000	−.3971385	−.1409511
mothsmo	−.1372306	.1846806	−0.74	0.457	−.4991979	.2247367
fathsmo	.0577346	.1023463	0.56	0.573	−.1428604	.2583296
bothsmo	−.0498802	.1134195	−0.44	0.660	−.2721783	.172418
alpa	−.1738161	.0273374	−6.36	0.000	−.2273963	−.1202358
alma	−.1094086	.0310003	−3.53	0.000	−.1701681	−.0486491
_cons	1.201467	1.136017	1.06	0.290	−1.025085	3.428018
nobese						
sc12	.1914998	.0690003	2.78	0.006	.0562617	.3267378
sc45	.0166942	.0664938	0.25	0.802	−.1136313	.1470197
lhqdg	.1064761	.1246465	0.85	0.393	−.1378265	.3507786
lhqhndA	−.0436436	.1162962	−0.38	0.707	−.27158	.1842927
lhqnone	−.0469588	.0963434	−0.49	0.626	−.2357884	.1418707

lhqoth	.0490445	.157629	0.31	0.756	−.2599028	.3579917
widow	−.2336653	.0895713	−2.61	0.009	−.4092219	−.0581088
divorce	−.1134252	.1391429	−0.82	0.415	−.3861403	.1592899
seprd	−.0557537	.2144662	−0.26	0.795	−.4760998	.3645923
single	−.1059766	.1264336	−0.84	0.402	−.3537818	.1418287
part	.0818783	.0946629	0.86	0.387	−.1036576	.2674141
unemp	−.138706	.1586382	−0.87	0.382	−.4496311	.1722191
sick	.0826019	.1662652	0.50	0.619	−.2432719	.4084758
retd	−.0686484	.1018029	−0.67	0.500	−.2681783	.1308816
keephse	−.1666617	.0982089	−1.70	0.090	−.3591477	.0258242
wkshft1	−.2303289	.1155888	−1.99	0.046	−.4568787	−.0037791
wales	−.2716858	.1213324	−2.24	0.025	−.5094929	−.0338787
north	.0204304	.1243057	0.16	0.869	−.2232042	.264065
nwest	−.0409076	.0998111	−0.41	0.682	−.2365338	.1547186
yorks	−.0557269	.1114449	−0.50	0.617	−.274155	.1627011
wmids	−.1036886	.1128301	−0.92	0.358	−.3248315	.1174543
emids	−.1857982	.1134226	−1.64	0.101	−.4081024	.036506
anglia	−.1859221	.1482363	−1.25	0.210	−.4764599	.1046157
swest	−.1555332	.1108586	−1.40	0.161	−.3728121	.0617457
london	.0850325	.1168399	0.73	0.467	−.1439694	.3140344
scot	−.2534751	.104264	−2.43	0.015	−.4578288	−.0491213
rural	.0890678	.0768817	1.16	0.247	−.0616176	.2397532
suburb	.2045193	.0622956	3.28	0.001	.0824221	.3266165
ethwheur	.1941513	.1807067	1.07	0.283	−.1600272	.5483298
housown	−.0552396	.1524866	−0.36	0.717	−.3541078	.2436287
hou	−.0266585	.0293081	−0.91	0.363	−.0841014	.0307843
height	.0040495	.0101182	0.40	0.689	−.0157819	.0238808
male	.4647972	.0833225	5.58	0.000	.301488	.6281064
age	−.0825432	.0246806	−3.34	0.001	−.1309162	−.0341701
age2	.0710613	.0201925	3.52	0.000	.0314848	.1106378
smother	−.0868442	.0594113	−1.46	0.144	−.2032881	.0295997
mothsmo	−.0927327	.166302	−0.56	0.577	−.4186787	.2332133
fathsmo	−.0543686	.0819079	−0.66	0.507	−.2149051	.1061679
bothsmo	.0829589	.0984885	0.84	0.400	−.110075	.2759928
alpa	−.0142073	.0229297	−0.62	0.536	−.0591486	.0307341
alma	.0706503	.0295389	2.39	0.017	.012755	.1285455
_cons	2.868913	1.03504	2.77	0.006	.8402722	4.897555
exercise						
sc12	−.0146504	.0568046	−0.26	0.796	−.1259854	.0966846
sc45	−.1662555	.0608175	−2.73	0.006	−.2854557	−.0470554

lhqdg	−.0257673	.0956648	−0.27	0.788	−.2132668	.1617322
lhqhndA	−.0334182	.0931971	−0.36	0.720	−.2160812	.1492448
lhqnone	−.2785791	.0787864	−3.54	0.000	−.4329975	−.1241606
lhqoth	−.0626586	.1235291	−0.51	0.612	−.3047712	.1794541
widow	−.0181257	.0847921	−0.21	0.831	−.1843152	.1480638
divorce	.2203262	.118629	1.86	0.063	−.0121824	.4528347
seprd	.1957486	.1759668	1.11	0.266	−.14914	.5406372
single	−.2916037	.1158734	−2.52	0.012	−.5187113	−.0644961
part	.1642691	.0783396	2.10	0.036	.0107264	.3178118
unemp	−.0515336	.132982	−0.39	0.698	−.3121734	.2091062
sick	−.6682178	.1581518	−4.23	0.000	−.9781896	−.3582459
retd	.0764447	.0865813	0.88	0.377	−.0932516	.246141
keephse	−.0746989	.0883851	−0.85	0.398	−.2479306	.0985328
wkshft1	−.181578	.0989575	−1.83	0.067	−.3755312	.0123751
wales	−.2238544	.1101651	−2.03	0.042	−.4397741	−.0079347
north	−.0728679	.1053162	−0.69	0.489	−.2792839	.133548
nwest	−.0551384	.0819348	−0.67	0.501	−.2157276	.1054508
yorks	−.02147	.0926616	−0.23	0.817	−.2030833	.1601433
wmids	−.272688	.0998771	−2.73	0.006	−.4684434	−.0769325
emids	−.0395094	.095307	−0.41	0.678	−.2263076	.1472889
anglia	−.0979963	.123006	−0.80	0.426	−.3390836	.1430911
swest	−.0417677	.0923038	−0.45	0.651	−.2226797	.1391444
london	.009507	.0923367	0.10	0.918	−.1714695	.1904835
scot	.0300929	.0900772	0.33	0.738	−.1464552	.206641
rural	.181503	.0676101	2.68	0.007	.0489897	.3140163
suburb	.1760586	.0545644	3.23	0.001	.0691143	.2830029
ethwheur	.1662137	.167461	0.99	0.321	−.1620038	.4944312
housown	.08902	.1397672	0.64	0.524	−.1849187	.3629586
hou	−.0712546	.0249737	−2.85	0.004	−.1202021	−.022307
height	.0003718	.0085756	0.04	0.965	−.0164359	.0171796
male	.1169778	.0706731	1.66	0.098	−.0215389	.2554945
age	−.027788	.0214826	−1.29	0.196	−.0698931	.0143171
age2	−.0047289	.0181378	−0.26	0.794	−.0402784	.0308206
smother	−.0941089	.0510864	−1.84	0.065	−.1942364	.0060186
mothsmo	−.0699969	.1461676	−0.48	0.632	−.3564801	.2164864
fathsmo	−.0024569	.0735997	−0.03	0.973	−.1467097	.1417959
bothsmo	.0038062	.0859619	0.04	0.965	−.1646759	.1722884
alpa	.0245778	.0205275	1.20	0.231	−.0156555	.064811
alma	.0165127	.025591	0.65	0.519	−.0336447	.0666701
cons	1.173644	.8773399	1.34	0.181	−.5459107	2.893198

rho21	.2782589	.2027816	1.37	0.170	−.1439988	.6147954
rho31	.3338423	.0905224	3.69	0.000	.146412	.4981336
rho41	−.2810325	.1164327	−2.41	0.016	−.4903913	−.041007
rho51	.3807708	.1429991	2.66	0.008	.0730311	.6223064
rho61	.039922	.176569	0.23	0.821	−.2974121	.3683949
rho71	−.2225948	.1238243	−1.80	0.072	−.4476261	.0289504
rho81	−.0890828	.1450253	−0.61	0.539	−.3590872	.1946814
rho32	.1119385	.097271	1.15	0.250	−.0804829	.2963163
rho42	−.1653874	.1467188	−1.13	0.260	−.4321775	.1280235
rho52	.0590127	.1773867	0.33	0.739	−.2819556	.3867459
rho62	.0001972	.1695367	0.00	0.999	−.3203963	.3207502
rho72	−.3534705	.1466959	−2.41	0.016	−.6030809	−.0408108
rho82	−.0140823	.2129715	−0.07	0.947	−.4066429	.3828679
rho43	.2699361	.0287952	9.37	0.000	.2126276	.325394
rho53	.0305489	.0288502	1.06	0.290	−.0260339	.0869366
rho63	.1825216	.0372825	4.90	0.000	.1085698	.2544643
rho73	−.2274227	.0353233	−6.44	0.000	−.2954055	−.1571483
rho83	.0950842	.0305487	3.11	0.002	.0349375	.1545446
rho54	.1243894	.0284744	4.37	0.000	.0682449	.1797486
rho64	.2684004	.0364378	7.37	0.000	.1956239	.3382369
rho74	.0767356	.0350884	2.19	0.029	.0077072	.1450361
rho84	.0728358	.0301832	2.41	0.016	.0134908	.1316695
rho65	.0489765	.036749	1.33	0.183	−.02318	.1206254
rho75	.0397135	.0335134	1.19	0.236	−.0260484	.1051332
rho85	−.0165165	.0284419	−0.58	0.561	−.0721527	.0392221
rho76	−.0449336	.0488476	−0.92	0.358	−.1399721	.0509252
rho86	−.0781129	.037887	−2.06	0.039	−.1518029	−.0035593
rho87	.1267338	.0350813	3.61	0.000	.0574751	.1947786

Likelihood ratio test of rho21=rho31=rho41=rho51=rho61=rho71=
rho81=rho32=rho42=rho52=rho62=rho72=rho82=rho43=rho53=rho63
=rho73=rho83=rho54=rho64=rho74=rho84=rho65=rho75=rho85=
rho76=rho86=rho87=0: chi2 (28)=276.392 Prob>chi2=0.0000

The rhos are the estimated correlation coefficients. Stata reports the asymptotic z-test for significance. This can be used to test the null hypothesis of exogeneity of the dummy regressors (see Knapp and Seaks 1998). The rhos for death-nsmoker, death-breakfast and death-sleepgd are statistically significant at a 5% significance level. The variables nsmoker, breakfast and sleepgd are also statistically significant determinants of mortality risk.

To refine the specification of the structural model, we consider a restricted version of the system of equations where only those lifestyles that have statistically significant values of rho in the general model are treated as endogenous:

• mvprobit (death=$xvars) (sah=$xvars) (nsmoker=$xvars (breakfast=$xvars) (sleepgd=$xvars), dr(50)
  • scalar logLmvp1=e(11)
  • disp "logL=" logLmvp1
  • scalar q1=e(k)
  • disp "q=" q1
  • scalar AIC=−2*logLmvp1+2*q1
  • disp "AIC=" AIC

• scalar BIC=−2*logLmvp1+log (N) *q1
  • disp "BIG=" BIG

• mvprobit (death=$deq) (sah=$heq) (nsmoker=$leq) (breakfast=$leq) (sleepgd=$leq), dr(50)

• scalar logLmvp0=e (11)
  • disp "logL=" logLmvp0
  • scalar q0=e (k)
  • disp "q=" q0
  • scalar AIC=−2*logLmvp0+2*q0
  • disp "AIC=" AIC
  • scalar BIC=−2*logLmvp0+log (N) *q0
  • disp "BIG:" BIG

• scalar testLR=2* (logLmvp0− logLmvp1)
  • disp testLR
  • scalar qLR=q0−q1
  • disp "q="qLR
  • disp chi2tail (qLR, testLR)

Here we calculate AIC=−2log$L$+2q, BIC=−2log$L$+$q$log$N$ and LR= −2(log$L_{unrestr}$–logL_{restr}) directly because fitstat cannot be used after mvprobit:

	Mvprobit with exclusion restrictions	Mvprobit without exclusion restrictions
AIC (Akaike)	20250.393	20285.582
BIC (Schwarz)	21491.166	21700.060
LR-test	χ^2_{28}=20.812, p-value=0.833	

The model with exclusion restrictions is favoured by the penalized likelihood criteria, and Table 5.4 reports the estimation results for the restricted version of the model.

Table 5.4 Multivariate probit—5 equations

Multivariate probit (SML, # draws=50) Number of obs= 3655
Wald chi2 (190)= 2890.80
Log likelihood=−9925.1967 Prob>chi2= 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
death						
sah	−.555212	.3258079	−1.70	0.088	−1.193784	.0833598
nsmoker	−.8679029	.1733127	−5.01	0.000	−1.20759	−.5282162
breakfast	.4144938	.2111862	1.96	0.050	.0005765	.8284111
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sleepgd	−.6813081	.2456369	−2.77	0.006	−1.162748	−.1998686
alqprud	−.0980097	.0723045	−1.36	0.175	−.2397239	.0437045
nobese	−.128288	.068301	−1.88	0.060	−.2621554	.0055795
exercise	−.0598006	.0582237	−1.03	0.304	−.173917	.0543158
sc12	−.069394	.0666233	−1.04	0.298	−.1999733	.0611854
sc45	−.0362109	.0636009	−0.57	0.569	−.1608663	.0884446
lhqdg	.0455054	.1155414	0.39	0.694	−.1809517	.2719624
lhqhndA	−.0562326	.1126342	−0.50	0.618	−.2769915	.1645263
lhqnone	.0633762	.0963374	0.66	0.511	−.1254417	.2521941
lhqoth	−.0841963	.1419392	−0.59	0.553	−.362392	.1939994
part	.1635744	.0993117	1.65	0.100	−.0310729	.3582217
unemp	.259161	.1396522	1.86	0.063	−.0145524	.5328744
sick	.2844689	.2202218	1.29	0.196	−.147158	.7160958
retd	−.0175596	.0945018	−0.19	0.853	−.2027796	.1676605
keephse	.2416762	.1057531	2.29	0.022	.0344038	.4489485
wkshft1	−.3001088	.1227498	−2.44	0.014	−.540694	−.0595235
rural	−.1001046	.0723458	−1.38	0.166	−.2418997	.0416906
suburb	−.0819488	.056881	−1.44	0.150	−.1934335	.029536
ethwheur	.3486228	.2011344	1.73	0.083	−.0455934	.742839
height	.0100251	.0091558	1.09	0.274	−.00792	.0279701
male	.3867326	.0763836	5.06	0.000	.2370235	.5364417
age	.0041715	.0260172	0.16	0.873	−.0468213	.0551643
age2	.0580687	.0213458	2.72	0.007	.0162318	.0999057
_cons	−2.532718	1.052304	−2.41	0.016	−4.595195	−.4702401
sah						
nsmoker	.0541581	.1903608	0.28	0.776	−.3189422	.4272585
breakfast	.3136966	.2715386	1.16	0.248	−.2185093	.8459025

sleepgd	.0798917	.3085474	0.26	0.796	−.5248501	.6846335
alqprud	−.0213176	.0746725	−0.29	0.775	−.1676731	.1250379
nobese	.1798918	.0635411	2.83	0.005	.0553536	.30443
exercise	.2528123	.0523655	4.83	0.000	.1501779	.3554468
sc12	.1936118	.0605934	3.20	0.001	.0748509	.3123727
sc45	−.1386097	.0579648	−2.39	0.017	−.2522186	−.0250008
lhqdg	.1239439	.1094928	1.13	0.258	−.090658	.3385457
lhqhndA	.0625327	.1036666	0.60	0.546	−.1406502	.2657156
lhqnone	−.1416623	.0875578	−1.62	0.106	−.3132725	.0299479
lhqoth	−.0798618	.1334242	−0.60	0.549	−.3413684	.1816449
widow	−.1734025	.0808545	−2.14	0.032	−.3318744	−.0149305
divorce	−.1247036	.1280383	−0.97	0.330	−.375654	.1262467
seprd	−.0095553	.1874956	−0.05	0.959	−.3770399	.3579293
single	.0090727	.1107018	0.08	0.935	−.2078988	.2260442
part	−.0592906	.0925852	−0.64	0.522	−.2407542	.122173
unemp	−.1789102	.1394462	−1.28	0.199	−.4522196	.0943993
sick	−1.556393	.1556369	−10.00	0.000	−1.861435	−1.25135
retd	−.2483696	.0913266	−2.72	0.007	−.4273665	−.0693727
keephse	−.2719767	.0936332	−2.90	0.004	−.4554943	−.088459
wkshft1	.006568	.1095614	0.06	0.952	−.2081685	.2213044
wales	−.271296	.1079963	−2.51	0.012	−.4829647	−.0596272
north	−.0087236	.1144178	−0.08	0.939	−.2329783	.2155311

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nwest	−.2099994	.08893	−2.36	0.018	−.384299	−.0356997
yorks	−.0998301	.0955071	−1.05	0.296	−.2870206	.0873605
wmids	−.0885406	.0998093	−0.89	0.375	−.2841632	.1070821
emids	−.00594	.1019029	−0.06	0.954	−.2056661	.1937861
anglia	−.1470017	.1268513	−1.16	0.247	−.3956257	.1016223
swest	−.1641106	.0956084	−1.72	0.086	−.3514996	.0232784
london	−.0616437	.0973267	−0.63	0.526	−.2524005	.1291132
scot	−.2390245	.0980292	−2.44	0.015	−.4311582	−.0468908
rural	.1063557	.0695595	1.53	0.126	−.0299784	.2426898
suburb	.0462762	.0561926	0.82	0.410	−.0638593	.1564116
ethwheur	.4592241	.1671204	2.75	0.006	.1316741	.7867741
housown	−.1921471	.1436457	−1.34	0.181	−.4736876	.0893934
hou	.0067001	.025684	0.26	0.794	−.0436395	.0570398
height	−.004124	.0088125	−0.47	0.640	−.0213961	.0131481
male	−.0039642	.0720215	−0.06	0.956	−.1451237	.1371953
age	−.0543932	.021038	−2.59	0.010	−.0956268	−.0131595
age2	.0462171	.0170244	2.71	0.007	.0128498	.0795843

_cons	1.883837	.9426364	2.00	0.046	.0363038	3.731371
nsmoker						
sc12	.1416012	.0600606	2.36	0.018	.0238846	.2593178
sc45	−.0835763	.0587207	−1.42	0.155	−.1986667	.0315141
lhqdg	.1129779	.1077724	1.05	0.294	−.098252	.3242079
lhqhndA	−.043605	.101754	−0.43	0.668	−.2430391	.1558291
lhqnone	−.214882	.0843267	−2.55	0.011	−.3801593	−.0496046
lhqoth	−.3578956	.1281365	−2.79	0.005	−.6090386	−.1067527
widow	−.2245397	.0843489	−2.66	0.008	−.3898604	−.059219
divorce	−.3703485	.120424	−3.08	0.002	−.6063751	−.1343219
seprd	−.2328926	.1751123	−1.33	0.184	−.5761064	.1103212
single	−.1568838	.1104866	−1.42	0.156	−.3734336	.0596661
part	−.102717	.0825477	−1.24	0.213	−.2645074	.0590735
unemp	−.3793703	.1322996	−2.87	0.004	−.6386726	−.1200679
sick	−.2556087	.1313316	−1.95	0.052	−.513014	.0017965
retd	−.2718833	.0894435	−3.04	0.002	−.4471894	−.0965772
keephse	−.121107	.0909171	−1.33	0.183	−.2993013	.0570873
wkshft1	−.1731242	.0991053	−1.75	0.081	−.367367	.0211186
wales	−.2002112	.1097018	−1.83	0.068	−.4152229	.0148004
north	−.3405112	.1035536	−3.29	0.001	−.5434726	−.1375498
nwest	−.3308475	.083653	−3.95	0.000	−.4948044	−.1668907
yorks	−.248911	.0952821	−2.61	0.009	−.4356603	−.0621616
wmids	−.2490437	.0980996	−2.54	0.011	−.4413153	−.056772
emids	−.0845381	.1023996	−0.83	0.409	−.2852376	.1161615
anglia	−.0833717	.1289559	−0.65	0.518	−.3361205	.1693772
swest	−.0853712	.0985469	−0.87	0.386	−.2785196	.1077772
london	−.1758891	.095992	−1.83	0.067	−.3640299	.0122517
scot	−.374853	.0922277	−4.06	0.000	−.5556161	−.1940899
rural	.0921524	.0693399	1.33	0.184	−.0437512	.2280561
suburb	.0210477	.0544729	0.39	0.699	−.0857173	.1278126
ethwheur	.0753455	.1631549	0.46	0.644	−.2444322	.3951233
housown	−.1084921	.1369728	−0.79	0.428	−.3769539	.1599697

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hou	.0532938	.0249165	2.14	0.032	.0044583	.1021293
height	.0095698	.0088597	1.08	0.280	−.007795	.0269346
male	−.249922	.0724836	−3.45	0.001	−.3919873	−.1078568
age	−.0502586	.0225363	−2.23	0.026	−.0944289	−.0060882
age2	.0630879	.01907	3.31	0.001	.0257114	.1004645
smother	−.7091654	.0512027	−13.85	0.000	−.8095209	−.6088099
mothsmo	−.4417742	.1457058	−3.03	0.002	−.7273524	−.156196

fathsmo	−.1922431	.0774122	−2.48	0.013	−.3439682	−.040518
bothsmo	−.2818085	.0876812	−3.21	0.001	−.4536605	−.1099565
alpa	−.043779	.0206955	−2.12	0.034	−.0843414	−.0032166
alma	−.0463402	.025267	−1.83	0.067	−.0958627	.0031823
cons	1.7851	.9135391	1.95	0.051	−.0054042	3.575603
breakfast						
sc12	.0499809	.0589114	0.85	0.396	−.0654833	.1654451
sc45	−.0448473	.0586814	−0.76	0.445	−.1598607	.0701662
lhqdg	.1826743	.1057211	1.73	0.084	−.0245352	.3898838
lhqhndA	−.0495355	.100472	−0.49	0.622	−.2464571	.147386
lhqnone	−.2754265	.0839345	−3.28	0.001	−.4399352	−.1109179
lhqoth	−.2189821	.1290411	−1.70	0.090	−.471898	.0339338
widow	−.1868454	.0848061	−2.20	0.028	−.3530624	−.0206285
divorce	−.2691698	.1181515	−2.28	0.023	−.5007424	−.0375972
seprd	−.3353321	.1742317	−1.92	0.054	−.67682	.0061558
single	−.0944547	.1099412	−0.86	0.390	−.3099356	.1210261
part	.1623416	.0821021	1.98	0.048	.0014244	.3232587
unemp	−.2931751	.129775	−2.26	0.024	−.5475294	−.0388208
sick	−.204853	.1265374	−1.62	0.105	−.4528618	.0431558
retd	.238393	.0892332	2.67	0.008	.0634991	.413287
keephse	−.0480381	.0881979	−0.54	0.586	−.2209028	.1248265
wkshft1	−.1516937	.0962014	−1.58	0.115	−.3402451	.0368576
wales	.1168218	.1078568	1.08	0.279	−.0945736	.3282173
north	.1649127	.1052673	1.57	0.117	−.0414074	.3712327
nwest	.2388051	.085006	2.81	0.005	.0721964	.4054139
yorks	−.0039636	.0927529	−0.04	0.966	−.185756	.1778287
wmids	.0060845	.0942838	0.06	0.949	−.1787083	.1908772
emids	−.0676967	.0964809	−0.70	0.483	−.2567957	.1214023
anglia	−.0198607	.1263212	−0.16	0.875	−.2674457	.2277242
swest	.065202	.0948479	0.69	0.492	−.1206965	.2511005
london	.0725208	.0937508	0.77	0.439	−.1112273	.256269
scot	.2120209	.0934429	2.27	0.023	.0288762	.3951656
rural	.1206431	.0678656	1.78	0.075	−.0123711	.2536573
suburb	.168355	.0541207	3.11	0.002	.0622804	.2744296
ethwheur	.4738206	.1519701	3.12	0.002	.1759648	.7716765
housown	−.1365436	.1334448	−1.02	0.306	−.3980907	.1250035
hou	−.030189	.0242809	−1.24	0.214	−.0777786	.0174006
height	.0156052	.0087884	1.78	0.076	−.0016197	.0328301
male	−.1149309	.0720763	−1.59	0.111	−.2561978	.0263359
age	.0238927	.0209138	1.14	0.253	−.0170976	.0648829

age2	−.009359	.0174691	−0.54	0.592	−.0435978	.0248797
smother	−.3415561	.050473	−6.77	0.000	−.4404814	−.2426309

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mothsmo	−.1827553	.1406348	−1.30	0.194	−.4583944	.0928838
fathsmo	.0712881	.0730876	0.98	0.329	−.0719611	.2145372
bothsmo	−.1248441	.083808	−1.49	0.136	−.2891047	.0394165
alpa	−.0474575	.0205421	−2.31	0.021	−.0877193	−.0071956
alma	−.0199611	.0252468	−0.79	0.429	−.0694439	.0295218
_cons	−1.544273	.874241	−1.77	0.077	−3.257754	.1692076
sleepgd						
sc12	.032793	.0544461	0.60	0.547	−.0739194	.1395054
sc45	.0213785	.0552699	0.39	0.699	−.0869485	.1297054
lhqdg	.0390026	.095131	0.41	0.682	−.1474508	.225456
lhqhndA	−.047011	.0925043	−0.51	0.611	−.2283162	.1342942
lhqnone	−.0726345	.0777189	−0.93	0.350	−.2249607	.0796917
lhqoth	.051098	.1213844	0.42	0.674	−.186811	.289007
widow	−.1969409	.0740873	−2.66	0.008	−.3421494	−.0517324
divorce	−.2545591	.112186	−2.27	0.023	−.4744396	−.0346786
seprd	−.3321837	.1642614	−2.02	0.043	−.6541302	−.0102373
single	−.1906472	.0996762	−1.91	0.056	−.386009	.0047146
part	.3178433	.0775442	4.10	0.000	.1658595	.4698272
unemp	.2302699	.1308771	1.76	0.079	−.0262445	.4867842
sick	−.2066586	.1237944	−1.67	0.095	−.4492912	.0359741
retd	.1454294	.080581	1.80	0.071	−.0125065	.3033653
keephse	.1959706	.0853507	2.30	0.022	.0286864	.3632548
wkshft1	−.281163	.0943412	−2.98	0.003	−.4660683	−.0962577
wales	.0320296	.1006168	0.32	0.750	−.1651757	.2292348
north	−.1443791	.0966898	−1.49	0.135	−.3338876	.0451293
nwest	.1150067	.0780594	1.47	0.141	−.0379869	.2680004
yorks	.0041006	.0863314	0.05	0.962	−.1651059	.1733071
wmids	−.0722429	.0891916	−0.81	0.418	−.2470552	.1025694
emids	−.083891	.0899885	−0.93	0.351	−.2602653	.0924833
anglia	−.0938982	.1151095	−0.82	0.415	−.3195087	.1317124
swest	.0012373	.0870094	0.01	0.989	−.169298	.1717726
london	−.0352171	.0853488	−0.41	0.680	−.2024977	.1320635
scot	−.1622086	.0842142	−1.93	0.054	−.3272653	.0028481
rural	.0561372	.0636191	0.88	0.378	−.0685538	.1808283
suburb	−.0438039	.0505733	−0.87	0.386	−.1429257	.0553179
ethwheur	.3539658	.1534294	2.31	0.021	.0532498	.6546819
housown	−.5080411	.1276151	−3.98	0.000	−.7581621	−.2579201

hou	−.0192479	.0232914	−0.83	0.409	−.0648982	.0264025
height	.0027891	.0081095	0.34	0.731	−.0131052	.0186835
male	.0483035	.0655495	0.74	0.461	−.0801712	.1767783
age	−.024044	.0186612	−1.29	0.198	−.0606192	.0125312
age2	.0100192	.0151753	0.66	0.509	−.0197238	.0397622
smother	−.0394418	.0482656	−0.82	0.414	−.1340407	.0551571
mothsmo	.2518358	.137515	1.83	0.067	−.0176886	.5213602
fathsmo	.0866362	.0666436	1.30	0.194	−.0439828	.2172551
bothsmo	.0526225	.0784872	0.67	0.503	−.1012097	.2064547
alpa	−.0186419	.0187939	−0.99	0.321	−.0554772	.0181934
alma	.005515	.0235375	0.23	0.815	−.0406176	.0516475
cons	1.21268	.8077068	1.50	0.133	−.3703958	2.795757

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rho21	.2222766	.2217302	1.00	0.316	−.2270916	.5936089
rho31	.3147777	.0992946	3.17	0.002	.1093824	.4943929
rho41	−.248028	.1239105	−2.00	0.045	−.4715721	.0054691
rho51	.3727766	.1500215	2.48	0.013	.0501107	.6249797
rho32	.0770073	.1017965	0.76	0.449	−.1229228	.2709304
rho42	−.1659688	.148449	−1.12	0.264	−.4355409	.1309219
rho52	.0319025	.1878481	0.17	0.865	−.324472	.3803459
rho43	.2712345	.0287714	9.43	0.000	.2139688	.3266423
rho53	.0311553	.0288991	1.08	0.281	−.0255253	.0876362
rho54	.1223829	.0284995	4.29	0.000	.0661951	.1777969

Likelihood ratio test of rho21=rho31=rho41=rho51=rho32=rho42=
rho52=rho43=rho53=rho54=0: chi2 (10)=120.487
Prob>chi2=0.0000

Testing exogeneity is straightforward. We use the z-tests as reported in Table 5.4, which show that the null of exogeneity (H_0: ρ_{jk}=0) is rejected for the lifestyle variables that are included. We interpret this evidence as saying that unobservable factors that influence the probability of being a nonsmoker (eating breakfast or sleeping well) also influence the probability of dying. The correlation between the mortality equations and the error terms of the non-smoker and sleep-well equations is positive, meaning that unobserved factors that increase the probability of being a non-smoker and sleeping well, also increase the mortality risk. In the case of breakfast, the correlation coefficient has, as expected, a negative sign. This suggests that assuming lifestyles are exogenous generates downward biased estimates of the effects of these behaviours. Accounting for endogeneity of the regressors allows us to capture a statistically significant effect of unobserved factors both on the mortality risk and the probability of some of the health-related behaviours.

We can also calculate an LR-test that replicates the one computed by mvprobit:

• qui probit death $deq
  • scalar logL1=e (11)
  • qui probit sah $heq
  • scalar logL2=e (11)

• qui probit nsmoker $leq
  • scalar logL3=e (11)
  • qui probit breakfast $leq
  • scalar logL4=e (11)
  • qui probit sleepgd $leq
  • scalar logL5=e (11)

• scalar logL_restr=logL1+logL2+logL3+logL4+logL5
  • disp logL_restr
  • scalar testLR=2* (logLmvp1—logL_restr)
  • disp testLR
  • disp chi2tail (10, testLR)

The LR test is used to test exogeneity, comparing the log-likelihood of the multivariate probit model to the sum of the log-likelihoods of the marginal probit models, estimated separately. These should be equal in the case of independent errors across the marginal distributions, ergo the LR test compares an unrestricted model to a restricted one, considering the separate probit estimates as a multivariate probit in which all correlations are restricted to zero. The null is rejected, which confirms the previous findings on endogeneity.

We calculate the average partial effects (APE) after mvprobit using the programs meffcon for continuous variables and meffdum for dummy variables. The partial effects are calculated for each observation using the latent index (here the post-estimation command to be used is mvppred, with option xb) and the probability of a non-zero dependent variable (mvppred, with option pmarg) and then averaged across individuals. The Stata code to define the programs for the marginal and average effects is as follows:

```
capture program drop meffcon
   program define meffcon
   version 9
         args pred beta
         quietly{
         gen meffcon=(‘beta’)*(normden(‘pred’))
         }
         summarize meffcon
         drop meffcon
   end

capture program drop meffdum
   program define meffdum
   version 9
```

```
        args pmarg pred beta covar
        quietly {
        gen meffdum=(‘pmarg’−norm(‘pred’−‘beta’))
        replace meffdum= norm(‘pred’+‘beta’)-(‘pmarg’) if
(‘covar’)==0
        }
        summarize meffdum
        drop mef fdum
end
```

```
• mvppred xb
    • drop xb2 xb3
    • mvppred pmarg, pmarg
    • drop pmarg2 pmarg3
    • sum xb1 pmarg1
    • foreach x of global xcont{
        disp “‘x’”
        meffcon xb _b[‘x’]
        }
    • foreach x of global xdum{
        disp “‘x’”
        mef fdum pmarg xb _b[‘x’] ‘x’
        }
```

We compare the partial effects from the recursive model with the partial effects from univariate probit models for mortality, both including exogenous lifestyles and excluding them, in order to assess the advantages of estimating a model that controls for endogeneity. The way we calculate the partial effects is different from the post-estimation command dprobit, since the latter calculates the partial effects at specific regressor values.

```
• qui probit death $deq/*, nolog*/
    • predict xb, xb
    • predict pmarg, p
    • global xcont “height age age2”
    • global xdum “sah nsmoker breakfast sleepgd alqprud nobese exercise sc12 sc45
lhqdg lhqhndA lhqnone lhqoth part unemp sick retd keephse wkshft1 rural suburb
ethwheur male”
    • foreach x of global xcont{
        disp “‘x’”
        meffcon xb b[‘x’]
        }
    • foreach x of global xdum{
        disp “‘x’”
        mef fdum pmarg xb _b[‘x’] ‘x’
        }
```

```
• drop xb pmarg
• qui probit death $deqex/*, nolog*/
• predict xb, xb
• predict pmarg, p
• global xcont “height age age2”
• global xdum “sc12 sc45 lhqdg lhqhndA lhqnone lhqoth part unemp sick retd keephse
wkshf t1 rural suburb ethwheur male”
• foreach x of global xcont{
        disp “‘x’”
        meffcon xb _b[‘x’]
        }
• foreach x of global xdum{
        disp “‘x’”
        mef fdum pmarg xb _b[‘x’] ‘x’
        }
• drop xb pmarg
```

Table 5.5 Average partial effects from alternative models for mortality

	Probit without endogenous dummy		Probit with exogenous dummy		Multivariate probit	
	APE	S.D.	APE	S.D.	APE	S.D.
sah			−0.075	0.036	−0.143	0.064
nsmoker			−0.087	0.042	−0.220	0.096
Breakfast			−0.038	0.019	0.101	0.052
sleepgd			−0.019	0.010	−0.175	0.077
alqprud			−0.029	0.014	−0.024	0.012
nobese			−0.043	0.021	−0.032	0.016
exercise			−0.022	0.011	−0.015	0.007
sc12	−0.045	0.021	−0.031	0.016	−0.017	0.009
sc45	0.006	0.003	−0.004	0.002	−0.009	0.005
lhqdg	0.001	0.001	0.008	0.004	0.011	0.006
lhqhndA	−0.020	0.010	−0.020	0.010	−0.013	0.007
lhqnone	0.040	0.018	0.022	0.011	0.016	0.008
lhqoth	−0.005	0.002	−0.020	0.010	−0.020	0.011
part	0.033	0.015	0.036	0.018	0.040	0.020
unemp	0.102	0.043	0.072	0.034	0.065	0.032
sick	0.227	0.085	0.156	0.068	0.072	0.034
retd	0.023	0.010	0.016	0.008	−0.004	0.002
keephse	0.077	0.034	0.061	0.030	0.060	0.030
wkshft1	−0.047	0.023	−0.062	0.033	−0.072	0.038

rural	−0.053	0.025	−0.038	0.020	−0.025	0.013
suburb	−0.026	0.012	−0.018	0.009	−0.020	0.010
ethwheur	0.078	0.040	0.092	0.051	0.082	0.044
height	0.002	0.001	0.002	0.001	0.003	0.001
male	0.111	0.049	0.107	0.052	0.096	0.047
age	0.009	0.004	0.009	0.004	0.001	0.001
age2	0.009	0.004	0.010	0.005	0.014	0.007

The results are summarized in Table 5.5, which reports the average of the partial effects (APE) and standard deviations from summarize. Notice that standard deviation reflects heterogeneity across the point estimates for each individual in the sample (unlike the standard error, which reflects sampling variation around a particular point estimate).

Table 5.5 shows that including lifestyles and health even under the restrictive assumption of exogeneity has a strong impact on the APE of the socioeconomic variables. Controlling for endogeneity in the multivariate probit model we find that lifestyles have a high impact on the risk of mortality relative to socioeconomic characteristics. In particular, the estimated APE of nsmoker, breakfast and sleepgd are much higher than those of the socioeconomic variables. For example, the risk of mortality for a non-smoker is about 22% lower than for a current smoker.

5.5 OVERVIEW

In this case study we investigated the extent that differences in the risk of mortality depend on lifestyle and individual socioeconomic characteristics, focusing mainly on social class and educational differences in the sample.

We relate the risk of mortality to a set of observable and unobservable factors. Observable factors influencing mortality are perceived health, socioeconomic and demographic characteristics, ethnicity, type of area and individual health-related behaviours. Individuals' choices about their lifestyle may induce variations in health status and affect mortality. We assume that the relationship between the socioeconomic environment and mortality risk is mediated by lifestyles. In order to assess the impact of lifestyles, we estimate probit models and compare models without lifestyles and models which include them.

The main econometric issue arising here is unobservable individual heterogeneity and endogeneity of the discrete explanatory variables, which is corrected by estimating a multivariate probit model for a recursive system of equations for deaths, health and lifestyles. We find that lifestyles have a high impact on the risk of mortality relative to socioeconomic characteristics.

Part III
Survival data

6
Smoking and mortality

6.1 INTRODUCTION

This chapter is about the use of survival analysis in health economics. The aim is to give the reader an insight into the modelling of continuous-time duration data, the use of non-parametric and parametric procedures and estimation methods that are commonly employed to analyse survival times. We present an application of these techniques to smoking initiation and cessation and the hazard of mortality, using data from the Health and Lifestyle Survey (HALS). The analysis focuses on the socioeconomic gradient in smoking duration and survival probability and the impact of smoking behaviour on survival probability.

Smoking trends are often associated with socioeconomic inequalities. Systematic differences in tobacco consumption exist between individuals with high and low socioeconomic status. Statistics show that in the UK the highest prevalence of smoking is among the poorest group in the population. US and European studies show that men from lower socioeconomic groups have a higher risk of dying from smoking-related diseases then men from upper groups (Kunst *et al.* 2004). People who have experienced social and economic disadvantages during childhood, adolescence and adult life may run the greatest risk of becoming addicted to nicotine and smoking.

The epidemiological evidence from the 1950s to date suggests that tobacco smoking is responsible for about 30% of cancer deaths in developed countries and it also causes deaths from vascular, respiratory and other diseases (Vineis *et al.* 2004). A host of studies show that mortality patterns are likely to be affected by the proportion of persons who give up smoking, and tobacco-related diseases account for a large proportion of all-cause mortality in all Europeans countries (see e.g., Peto *et al.* 2005). Therefore a deeper investigation of the relationship between smoking behaviour and survival probability is necessary.

6.2 BASIC CONCEPTS OF SURVIVAL ANALYSIS

In health economics, as in other fields of economics, many variables indicate the time elapsed before an event occurs. Therefore these variables are in the form of a duration. Time to death, time to starting using a drug and time to quitting are typical examples. Survival time data give additional information relative to binary variables describing the occurrence of an event (death) or the choice of participation (starting or quitting).

In this chapter we focus on continuous time data assuming that the transition event may occur at any instant in time, while Chapter 7 covers discrete time models. In particular, we define the length of a spell for an individual in the sample as the realisation

of a continuous random variable, T, that has the following cumulative distribution function (cdf):

$F(t)=P(T\leq t)$

The cdf is known as the *failure function,* its complement is the *survivor function,* which indicates the probability of surviving up to a specific point in time t and can be defined as:

$S(t)=1-F(t)=P(T > t)$ where $0\leq S(t)\leq 1$

The probability of survival is equal to 1 at entry in the state of interest. The density function, which is the slope of the failure function, indicates the concentration of failure times along the time axis, and is expressed by:

$$f(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}$$

The hazard function is the instantaneous rate of failure per unit of time, conditional on individual survival up to that instant, and can be expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \cdot \frac{1}{1 - F(t)} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Integrating the hazard rate we obtain the *cumulative hazard function,* which sums up the hazard at each instant in time:

$$H(t) = \int_0^t h(u)du$$

Survival analysis constructs variables indicating the length of time a person stays in the state of interest. Usually respondents in samples are asked about the date of entry and exit from the state, as in the case of HALS.

Individuals are assumed to enter the state at time 0 and leave it at some time t, when the event of failure occurs. If entry and failure are observed, it will be possible to measure a complete spell. While, if only entry is observed and exit will eventually occur at some time T in the future, the spell will be incomplete. Such incomplete durations are known as *right-censored* spells, where censoring is at the time of observation, and we only know that the complete duration will be $T>t$. When the date of entry is not known we cannot measure the exact length of the spell and the survival time is said to be *left-censored.*

Depending on the state of interest, only those individuals who have survived for a minimum amount of time in the state are included in the sample or, putting the problem another way, only individuals who fail before the time of observation will not be included. Hence, the remaining observed survival times are said to be *left-truncated* (to describe left-truncation Stata refers to the concept of *delayed entry*). In the analysis we use, at most, one completed spell per individual. *Right-truncation* can be forced by the researcher who wants to restrict the population sample to those individuals who failed by the observation time, thus eliminating all longer survival times. More complete readings on survival analysis are available in Wooldridge (2002b) and Cameron and Trivedi (2005). For a more applied approach to survival analysis see also Jenkins (2004).

6.3 THE HALS DATA

The HALS data describe a representative sample of the British population as of 1984 and provide information about individual mortality by tracking each respondent on the NHS registers on a regular basis. The latest deaths data were released in June 2005. This allows us to investigate survival up to April 2005.

As shown in Chapter 5, the status of the respondent by the last update of the survey can be explored in Stata using the command:

• tab flagcode

current flagging status April 05	Freq.	Percent	Cum.
on file	6,248	69.40	69.40
not nhs regist.	85	0.94	70.34
deceased	2,431	27.00	97.35
rep. dead not id	1	0.01	97.36
embarked -abroad	42	0.47	97.82
no flag yet rec.	196	2.18	100.00
Total	9,003	100.00	

Up to April 2005, 97.8% of the original sample had been flagged and 27% of the respondents had died.

For the purpose of our analysis the target sample has been reduced according to item non-response in the variables of interest. Furthermore, the cumulative distribution of deathage (age at death) suggests restricting the analysis to individuals older than 40 years. This is shown by the command summarize reported below, where the option detail is used to produce various percentiles for deathage:

• summ deathage if deathage !=0, detail

age at death

Percentiles		Smallest		
1%	40.2	25.7		
5%	53.9	26.3		
10%	59.8	26.7	Obs	2105
25%	68.5	26.9	Sum of Wgt.	2105
50%	76.7		Mean	75.27515
		Largest	Std. Dev.	11.8917
75%	83.7	100.5		
90%	89	100.6	Variance	141.4124
95%	91.8	101.2	Skewness	−.7590558
99%	96.9	101.6	Kurtosis	3.875779

This shows that only 1% of the sample died before age 40 and that the average age at death if deathage !=0, that is for those who died, is 75.

The following command generates a global list of variables:

• global vars “birthmth birthday birthyr seenmth seenday seenyr age death deathage agestrt exfag exfagan regfag sc12 sc3 sc45 lhqdg lhqoth lhqnone lhqO lhqA lhqhnd rural married widow sepdiv single part unemp sick retd keephse wkshf t1 housown hou suburb mothsmo f athsmo bothsmo smother male”

These are then described:

• describe $vars

variable name	storage type	display format	value label	variable label
birthmth	byte	%8.0g		month of birth
birthday	byte	%8.0g		birth day of month
birthyr	byte	%8.0g		year of birth
seenmth	byte	%8.0g		SEENMTH
seenday	byte	%8.0g		SEENDAY
seenyr	byte	%8.0g		SEENYR
age	float	%9.0g		age at HALS1
death	float	%9.0g		
deathage	double	%10.0g		age at death
agestrt	float	%9.0g		age at starting smoking
exfag	byte	%9.0g		if ex-smoker

exfagan	byte	%4.0g	how long ago stopped smoking
regfag	byte	%9.0g	1 if smokes regularly at least one fag a day
sc12	float	%9.0g	1 if professional/student or managerial/intermediate
sc3	float	%9.0g	1 if skilled or armed service
sc45	float	%9.0g	1 if partly skilled, unskilled, unclass. or never occupied
lhqdg	byte	%9.0g	1 if University degree
lhqoth	byte	%9.0g	1 if other vocational/professional qualifications
lhqnone	byte	%9.0g	1 if no qualification
lhqO	byte	%9.0g	1 if O level/CSE
lhqA	byte	%9.0g	
lhqhnd	byte	%9.0g	
rural	byte	%8.0g	1 if lives in the countryside
married	byte	%8.0g	1 if married
widow	byte	%8.0g	1 if widow
sepdiv	float	%9.0g	1 if separated or divorced
single	byte	%8.0g	1 if single
part	byte	%8.0g	1 if part time worker
unemp	byte	%9.0g	1 if the individual unemployed
sick	byte	%9.0g	1 if absent from work due to sickness
retd	byte	%8.0g	1 if retired
keephse	byte	%8.0g	1 if housekeeper
wkshft1	float	%9.0g	1 if shift worker
housown	byte	%9.0g	1 if own or rent house
hou	byte	%9.0g	number of other people in the house
suburb	byte	%8.0g	1 if lives in the suburbs of the city
mothsmo	float	%9.0g	1 if only mother smoked
fathsmo	float	%9.0g	1 if only father smoked

bothsmo	float	%9.0g	1 if both parents smoked
smother	byte	%4.0g	1 if anyone else in house smoked
male	byte	%9.0g	1 if male

6.4 SURVIVAL DATA IN HALS

The HALS questionnaire is designed to provide comprehensive information about risky behaviours. In particular, the survey data contain retrospective information on smoking. The age at the onset of smoking is known as well as whether they are regular smokers or they stopped completely, and how long ago. The self-reported variables agestrt, exfag, exfagan, regfag are used to derive two time variables which can be used to study the hazard of starting smoking and the hazard of quitting smoking. This follows Forster and Jones (2001), who use the HALS data to investigate the role of tobacco taxes in starting and quitting smoking.

Smoking initiation

We define starting as the number of years elapsed before someone starts smoking. First, we adjust the variable agestrt to have the true age at starting smoking:

• gen agestart=.
  • replace agestart=agestrt*10

We eliminate individuals who claimed to be current smokers but whose age at starting was zero and we generate the binary indicator start that indicates whether an individual started smoking at some point in their life prior to HALS:

• drop if regfag==1 & agestart==0
  • gen start=.
  • replace start=1 if agestart>0
  • replace start=0 if start==.
  • label variable start “eversmoker”

The time variable starting measures a complete duration if an individual had started smoking and an incomplete, or censored, duration if they had not:

• gen starting=agestart if start==1
  • replace starting=age if start==0
  • label variable starting “number of years non-smoking”

For those who started smoking, starting is equal to the age at starting (agestart), and it is censored at the age at the time of the interview for those who had not started by then.

Smoking cessation

The variable exfagan provides information about how long ago an ex-smoker stopped smoking. This can be used to build a time variable indicating the number of years a person smoked. In order to derive the survival time variable for smoking we also need to exploit the exact information available about the onset of smoking. First, we generate the binary indicator quit that takes the value of one if a smoker had stopped smoking completely and zero otherwise:

```
• gen quit=.
    • replace quit=1 if exfag==1&exfagan<98
    • replace quit=0 if regfag==1
```

The definition of the dummy variable quit requires consistent information from the variables regfag, exfag and exfagan, in particular only values of exfagan smaller than 97 can be used, because higher values identify a missing record.

We generate each individual's date of interview using the mdy (m, d, y) date function, which returns the elapsed date corresponding to the numeric arguments of month, day and year, and we use the command format, which specifies that the variable will be displayed in default numeric elapsed date format:

```
• gen seenmdy=mdy(seenmth,seenday,seenyr+1900)
    • format seenmdy %d
```

We then use the variables birthmth, birthday, birthyr to generate the exact date of birth.

```
• generate birthmdy=mdy (birthmth, birthday, birthyr +1800) if birthyr>=87
    • format birthmdy %d
    • replace birthmdy=mdy (birthmth, birthday, birthyr +1900) if birthyr<87
    • format birthmdy %d
```

Cross-checking with the variable age in the original sample we notice that there cannot be individuals born before 1887 or after 1986. For this reason we use the if qualifier, which allows us to recover the true year of birth. The new variable birthmdy can be listed with the command list, which is used here to display the first ten observations in the data:

```
• list birthmdy in 1/10
```

	birthmdy
1.	18jun1930
2.	29mar1932
3.	25aug1918
4.	03mar1904
5.	07may1909
6.	11oct1937
7.	02jul1922

8.	28mar1932
9.	06feb1913
10.	31jul1922

We can now define the date of the onset of smoking:

• summ agestart if start==1, d
  • gen startmdy=birthmdy+(agestart*365.25) if start==1
  • replace startmdy=birthmdy+(17*365.25) if start==0
  • format startmdy %d

For those who did not start smoking before the survey, the date of starting is rescaled using the median age at starting, 17 years old, that we find by summarizing agestart, which is the age at which individuals are potentially at risk of starting:

• gen strtyear=year(startmdy)
  • summ strtyear
  • gen year=strtyear−1904
  • gen lnyear=ln (year)

We recover the year of starting using the function year () and, for the purpose of including the variable as a control variable in the econometric model, we rescale it relative to the earliest year, 1904, and take the logarithm.

We can now generate the time variable sm_years according to the smoking status of the respondent:

• gen sm_years=(seenmdy−startmdy)/365.25 if quit==0
  • replace sm_years=(seenmdy−startmdy− (exfagan*365.25))/365.25 if quit==1
  • gen sm_years2=round(sm_years, 1)
  • drop sm_years
  • rename sm_years2 sm_years

For current smokers, i.e. quit==0, smoking duration is censored at the time of the interview; for quitters, i.e. quit==1, the true duration is recovered using exfagan.

Lifespan

The HALS data give us the scope to investigate the hazard of death. A complete duration is observed for those who died before the follow-up period, while an incomplete duration is associated with individuals who were still alive in April 2005. We first generate the age at censoring and use that variable to create the time variable lifespan:

• gen ageATcens=(mdy(4,1,2005)−birthmdy)/365.25 if death==0
  • gen lifespan=.
  • replace lifespan=deathage if death==1
  • replace lifespan=ageATcens if death==0

• label var lifespan “survival time: censoring at April 2005”

The variable lifespan is simply equal to the age at death, deathage, for those who died, and to the age at censoring, ageATcens, for those who are still alive. Our time variable lifespan assumes that the measure of duration begins at birth and measures the full lifespan. As an alternative one could assume that individuals enter the initial state only when they participate in the survey process, so that the entry date would be the seen date at HALS (see Cheung 2000). The advantage of defining lifespan in the way that we do is that we are able to measure length of survival from birth, conditional on survival up to the time of the survey, in which case the distribution of the survival time is said to be *left-truncated.*

Data cleaning

Having created the relevant time variables for the analysis, we test the data for consistency and decide to clean the dataset, dropping those individual who report negative or nil value for sm_years. We also drop those who are ex-smokers but either do not have a record for exfagan or do not report their age of starting smoking:

• list serno sm_years regfag exfag exfagan birthmdy agestart start startmdy
  • count if sm_years<=0
  • count if exfag==1&exfagan==.
  • count if exfag==1 & agestart==0
  • drop if sm_years<=0
  • drop if exfag==1&exfagan==.
  • drop if exfag==1 & agestart==0

6.5 DESCRIPTIVE STATISTICS

The following command produces summary statistics:

• summarize $vars

Variable	Obs	Mean	Std. Dev.	Min	Max
birthmth	4646	6.459105	3.434446	1	12
birthday	4646	15.79488	8.817068	1	31
birthyr	4646	27.54176	13.22798	0	99
seenmth	4646	6.440594	3.487919	1	12
seenday	4646	12.77099	9.544777	1	31
seenyr	4646	84.64507	.4785433	84	85
age	4646	58.04567	11.75395	40	96.8
death	4646	.4276797	.4947954	0	1
deathage	4646	32.75956	38.46887	0	101.6
agestrt	4646	1.126862	1.004295	0	7
exfag	4646	.3138183	.4640935	0	1

exfagan	1458	15.07407	11.89164	0	70
regfag	4646	.3105898	.4627849	0	1
sc12	4646	.2998278	.4582317	0	1
sc3	4646	.4739561	.499375	0	1
sc45	4646	.2262161	.4184257	0	1
lhqdg	4646	.1168747	.3213055	0	1
lhqoth	4646	.0501507	.2182793	0	1
lhqnone	4646	.6254843	.4840497	0	1
lhqO	4646	.0910461	.2877056	0	1
lhqA	4646	.0357297	.1856353	0	1
lhqhndA	4646	.1164443	.3207914	0	1
rural	4646	.2139475	.4101343	0	1
married	4646	.75226	.4317465	0	1
widow	4646	.1287129	.3349179	0	1
sepdiv	4646	.0555316	.2290397	0	1
single	4646	.0634955	.2438783	0	1
part	4646	.1289281	.3351564	0	1
unemp	4646	.0312096	.1739026	0	1
sick	4646	.0327163	.1779123	0	1
retd	4646	.3555747	.4787386	0	1
keephse	4646	.097288	.2963814	0	1
wkshft1	4646	.057684	.2331701	0	1
housown	4646	.9692208	.3030518	0	9
hou	4646	1.602454	1.25266	0	9

Variable	Obs	Mean	Std. Dev.	Min	Max
suburb	4646	.4623332	.4986329	0	1
mothsmo	4646	.0456307	.2087052	0	1
fathsmo	4646	.5798536	.4936353	0	1
bothsmo	4646	.2277228	.4194079	0	1
smother	4646	.3452432	.4754987	0	1
male	4646	.4543693	.4979671	0	1
start	4646	.6244081	.4843275	0	1
quit	2901	.5025853	.5000795	0	1
starting	4646	33.18405	21.54336	4	96.8
sm_years	2901	32.23716	14.00826	1	72
lifespan	4646	73.99948	9.62414	41.5	110.976

Our sample consists of 4,646 observations; 43% of the respondents had died by April 2005 and the mean lifespan is 74 years. Those who started smoking at some point in their life account for the 62% of the sample among whom about 50% had stopped smoking at

the time of HALS. On average, smokers (current and ex-smokers) in the sample smoked for 32 years. For some variables we need to restrict the sample according to death to calculate sample means.

• summ age if death==0

Variable	Obs	Mean	Std. Dev.	Min	Max
age	2659	51.98514	8.669153	40	90.6

• summ deathage if death==1, d

age at death

	Percentiles	Smallest		
1%	52	41.5		
5%	58.8	43.9		
10%	62.3	44.4	Obs	1987
25%	69.6	45.4	Sum of Wgt.	1987
50%	77.4		Mean	76.55702
		Largest	Std. Dev.	10.22046
75%	84.1	100.5		
90%	89.2	100.6	Variance	104.4577
95%	92	101.2	Skewness	−.2947964
99%	97.1	101.6	Kurtosis	2.759138

The mean age is about 52 and half of the respondents died before age 77. The mean age at death is around 76.

6.6 DURATION MODELS

Smoking initiation

We start investigating the onset of smoking using the binary choice start and the duration variable starting:

• tab start

eversmoker	Freq.	Percent	Cum.
0	1,745	37.56	37.56
1	2,901	62.44	100.00
Total	4,646	100.00	

• sum starting, d

number of years non-smoking

Percentiles		Smallest		
1%	8	4		
5%	12	4		
10%	14	5	Obs	4646
25%	16	5	Sum of Wgt.	4646
50%	21		Mean	33.18405
		Largest	Std. Dev.	21.54336
75%	50.6	92.7		
90%	67.1	93.9	Variance	464.1165
95%	74.2	94.9	Skewness	.7810813
99%	83.1	96.8	Kurtosis	2.243695

2,901 individuals in our sample started smoking and, on average, the length of time spent non smoking is about 33 years, with half of the sample surviving for at least 21 years. Notice that here we need to consider both starters and never-smokers, which implies that survival times longer than the completed durations are included.

Stata has a built-in command that allows us to declare that a variable contains survival time data. Using stset we check the consistency of the time data and ensure that they make sense:

• stset starting, failure (start)/*id(serno)*/

The command stset requires us to specify a few options that help to identify the duration of interest. We specify failure (start) because the duration starting is complete for those who start smoking or, in other words, the onset of smoking represents the failure event. The option id (serno) can be specified if there are repeated observations for each individual, as there would be with time varying covariates. In our case this option does not affect stset because we only have one record per individual. Stata provides the following output:

failure event:	start !=0 & start <.
obs. time interval	(0, starting]
exit on or before:	failure
4646	total obs.
0	exclusions
4646	obs. remaining, representing
2901	failures in single record/single failure data

154173.1	total analysis time at risk, at risk from t=	0
	earliest observed entry t=	0
	last observed exit t=	96.8

The full sample size is used because no exclusion has been specified using the if qualifier.

We can explore survival time data, summarizing both sample and individual survival time:

• stsum

failure _d: start
analysis time _t: starting

	time at risk	incidence rate	no. of subjects	|------Survival time-----| 25%	50%	75%
total	154173.1	.0188165	4646	16	21	

This shows total time at risk and the incidence rate calculated as the ratio between the number of failures and total time at risk (2,901/154,173.1).

• stdes

failure_d: start
analysis time_t: starting

Category	total	|------per subject-----| mean	min	median	max
no. of subjects	4646				
no. of records	4646	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		33.18405	4	21	96.8
subjects with gap	0				
time on gap if gap	0				
time at risk	154173.1	33.18405	4	21	96.8
failures	2901	.6244081	0	1	1

This shows that subjects enter the state at time 0 (i.e., there is no left-censoring) and that data contain a record for each subject at risk. Mean, median, minimum and maximum values of time at risk are reported.

The command sts graph produces graphs of the estimated failure, survivor, hazard and cumulative hazard functions obtained with nonparametric procedures. The functions *f(t), S(t)* and *h(t)* are estimated using the Kaplan-Meier (or product-limit) estimator. The observation period is divided into *k* survival times $t_1<t_2<\ldots<t_j<\ldots t_k$ such that the

beginning of each survival time corresponds to the previous failure. If there is no censoring, the empirical survivor function is:

$$\hat{S(t)} = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

Where n_j the number of persons at risk of making a transition and d_j is the number of persons for which exit is observed. Therefore, the estimated survivor function is the product for each time j of the ratio between those who survive and the total number of persons at risk; it has the shape of a step function with origin in $t=0$ and at each t_j the height is equal to $S(t_j)$. From the estimated survivor function one can derive the estimated failure function and the integrated hazard function.

The graph for sts graph produces a smooth hazard function (usually Kernel smoothers are used) and the option na allows one to graph the estimated Nelson-Aalen cumulative hazard function, which behaves better than the Kaplan-Meier estimator in small samples:

$$\hat{H(t)} = \sum_{j=1}^{k} \frac{d_j}{n_j}$$

We use the following commands:

• sts graph, title (“Kaplan-Meier Survivor-starting”) saving(KMsurv_start, replace)

• sts graph, hazard title (“Kaplan-Meier Hazard-starting”) saving(KMhazstart, replace)

• sts graph, na title (“Nelson-Aalen cumulative Hazard function-starting”)saving(NAcumhazstart, replace)

• gr combine “KMsurv_start” “KMhazstart” “NAcumhazstart”, saving(startingNP, replace)

Figure 6.1 shows the decreasing pattern of the survivor function, with survival diminishing faster from duration of 15 years to duration of 20 years and then less than proportionally. The hazard of starting smoking increases up to age 20 and then falls close to 0 for duration longer than 35 years.

The estimation of a parametric model for the onset of smoking requires some tests for the distribution that could best be used to represent the duration variable starting. We use both the cumulative Cox-Snell residuals and information criteria to discriminate among distributions and to choose the best fitting distribution. In particular we compare the exponential, Weibull, log-normal and log-logistic distributions.

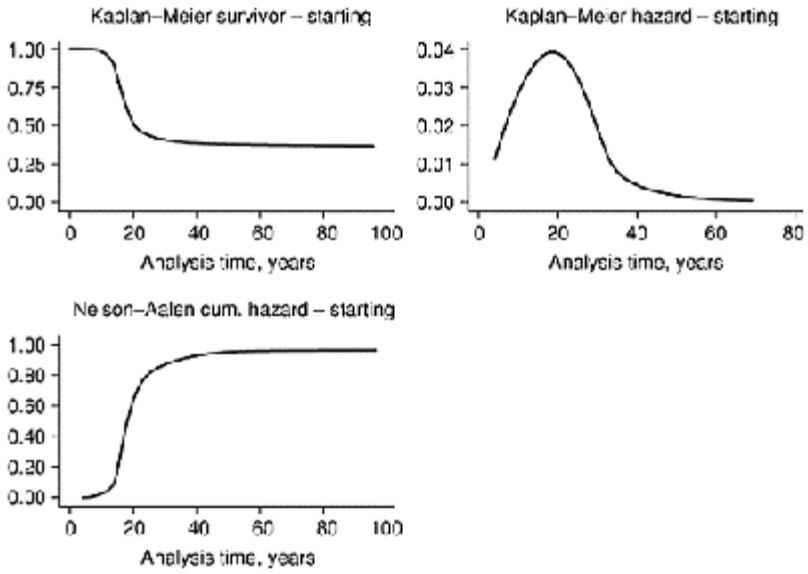

Figure 6.1 Non-parametric functions for smoking initiation.

The exponential is the most basic model, because it describes a flat hazard function, that is the hazard function is constant over time, and requires no additional parameter to be estimated (the 'shape parameter' is set equal to one). The Weibull has a more general form of the hazard function, which can be monotonically increasing or decreasing depending on the shape parameter. If the shape parameter equals one, then the Weibull reduces to the exponential distribution. The log-logistic and the log-normal distributions represent the logarithm of time using a logistic and a normal distribution, respectively. They tend to produce similar results: the shape parameter can describe either a monotonically decreasing hazard rate or a first increasing and then decreasing hazard rate.

We define a list of variables that we want to include as regressors. We want our duration model to be a flexible function of age, hence age is expressed in logarithms in order to have the elasticity of the hazard with respect to age. We also assume that the onset of smoking depends on socioeconomic variables (social class and education), the environment, and smoking behaviour of the household. Social class and education level at the observation time are assumed to reflect past socioeconomic characteristics.

- gen lnage=ln (age)
- global xstart "sc12 sc45 lhqdg lhqhndA lhqnone lhqoth rural suburb mothsmo fathsmo bothsmo smother male lnage"

We use the option clear to ask Stata to forget the st markers generated with the previous stset command and then stset the data again:

/*Exponential*/

- stset, clear
- qui stset starting, failure (start) id (serno)
- qui streg $xstart, d (exp) time
- *estat ic
- estimates store exp
- predict double cs, csnell
- stset, clear
- qui stset cs, failure (start)
- qui sts gen km=s
- qui gen double H=−ln (km)
- qui line H cs cs, sort title (“Exponential”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“ExpCSstart”, replace)

/*Weibull*/

- stset, clear
- qui stset starting, failure (start)
- qui streg $xstart, d (w) time
- estimates store weibull
- predict double cs2, csnell
- stset, clear
- qui stset cs2, failure (start)
- qui sts gen km2=s
- qui gen double H2=−ln (km2)
- qui line H2 cs2 cs2, sort title (“Weibull”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“WeibCSstart”, replace)

/*Log-Normal*/

- stset, clear
- qui stset starting, failure (start)
- qui streg $xstart, d (lognormal)
- estimates store logN
- predict double cs3, csnell
- stset, clear
- qui stset cs3, failure (start)
- qui sts gen km3=s
- qui gen double H3=−ln (km3)
- qui line H3 cs3 cs3, sort title (“logNormal”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“LogNormalCSstart”, replace)

/*Log-Logistic*/

- stset, clear
- qui stset starting, failure (start)

• qui streg $xstart, d (loglogistic)
• estimates store logL
• predict double cs4, csnell
• stset, clear
• qui stset cs4, failure (start)
• qui sts gen km4=s
• qui gen double H4=−ln (km4)
• qui line H4 cs4 cs4, sort title (“logLogistic”) ylabel (0 (0.5) 3) legend (off) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“LogLogisticCSstart”, replace)

• drop cs km H cs2 km2 H2 cs3 km3 H3 cs4 km4 H4

• gr combine “ExpCSstart” “WeibCSstart” “LogNormalCSstart” “LogLogisticCSstart”, imargin (1 10 1 10) graphregion (margin (1=2 r=2)) saving(“CoxSnell_start”, replace)

• estimates stats exp weibull logN logL
• drop_est*

For each distribution we quietly estimate the regression model by maximum likelihood using streg. The option d( ) is used to specify the distribution. All models are estimated in the accelerated failure time (AFT) metric specifying the option time for the models that do not have *a priori* an AFT parameterization. The csnell option in predict generates the Cox-Snell residuals. For observation j at time t_j the residuals are defined as

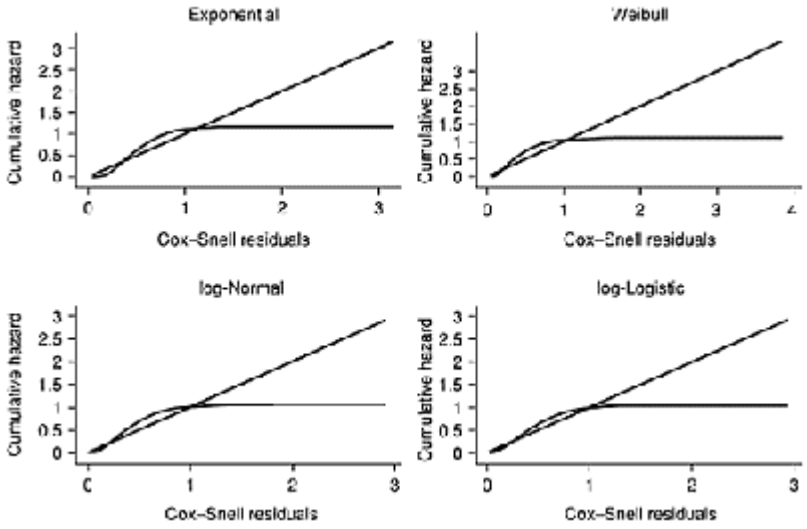

Figure 6.2 Cox-Snell residuals test—smoking initiation.

the cumulative hazard from the fitted model, $CS_j = -\ln \hat{S}_j(t_j)$, and are saved in cs. If the model fits the data the residuals are distributed as a standard exponential distribution with ancillary parameter equal to one. To verify this it is sufficient to estimate the Kaplan-Meier or the Nelson-Aalen cumulative hazard. We save the Kaplan-Meier survival estimate in km and the cumulative hazard in H. Then H is plotted against cs. A comparison between the four graphs shows that the distributions that fit the data a little better are the log-normal and the log-logistic (Figure 6.2).

We use the post-estimation command estimates store followed by the name of the model to store the statistics of the fitted model. The command estat ic could be used as well after each regression model but it does not allow us to store results. This command produces the information criteria AIC and BIC. Table 6.1 summarizes the results.

Table 6.1 Information criteria—smoking initiation

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
exp	4646	−6175.839	−5802.74	15	11635.48	11732.14
weibull	4646	−6152.564	−5712.719	16	11457.44	11560.54
logN	4646	−5588.572	−5208.542	16	10449.08	10552.18
logL	4646	−5695.691	−5285.974	16	10603.95	10707.05

The information criteria confirm that the models with the best fit are the log-normal and the log-logistic. We present results from the estimation of the log-logistic model, which has been shown to best fit the HALS data in Forster and Jones (2001). In the log-logistic model the hazard, the survival and the density functions are as follows:

$$h(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{\varphi_i^{1/\gamma} t_i^{(1/\gamma - 1)}}{\gamma[1 + (\varphi_i t_i)^{1/\gamma}]}$$

$$S(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = [1 + (\varphi_i t_i)^{1/\gamma}]^{-1}$$

$$f(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{\varphi_i^{1/\gamma} t_i^{(1/\gamma - 1)}}{\gamma[1 + (\varphi_i t_i)^{1/\gamma}]^2}$$

Where $\varphi_i = exp(-\boldsymbol{x}_i\boldsymbol{\beta})$ is a non-negative function that depends on observed characteristics and whose shape depends on the ancillary parameter γ. If $\gamma \geq 1$ the hazard rate is monotonically increasing, if $\gamma < 1$ then the hazard first rises and then decreases monotonically.

For a duration model with right censoring, the likelihood for individual *i* is the expression:

$$ln(t_i) = \boldsymbol{x}_i\boldsymbol{\beta}^* + \alpha u_i,$$

Where d_i is the failure, or censoring, indicator. Individuals who start smoking (i.e., they fail) have a complete spell for starting, and those who never start have a censored spell.

Hence, in this case, if start==1 the likelihood function is the duration density function; if start==0 the likelihood is the survivor function. For the log-logistic distribution the likelihood function becomes:

$$L_i = \left(\frac{\varphi_i^{1/\gamma} t_i^{(1/\gamma - 1)}}{\gamma [1 + (\varphi_i t_i)^{1/\gamma}]^2} \right)^{start} \cdot ([1 + (\varphi_i t_i)^{1/\gamma}]^{-1})^{(1-start)}$$

The log-logistic model has an AFT metric. Hence, the model can be written as $ln(t_i)=x_i\beta^*+\alpha u_i$, where α is the inverse of the ancillary parameter and u_i is an error term. The AFT metric assumes a linear relationship between the log of survival time t_i and characteristics $\boldsymbol{x_i}$.

We estimate the model with streg. There is no need to specify the option time (Table 6.2):

• stset, clear
  • qui stset starting, failure (start) id(serno)
  • streg $xstart, d(loglogistic) nolog

Table 6.2 Smoking initiation—log-logistic distribution (AFT)—coefficients

Log-logistic regression--accelerated failure-time form

No. of subjects=	4646	Number of obs=	4646
No. of failures=	2901		
Time at risk=	154173.1		
		LR chi2 (14)=	819.43
Log likelihood=	−5285.9735	Prob>chi2=	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sc12	.0997876	.0363994	2.74	0.006	.0284461	.1711292
sc45	−.079418	.0348648	−2.28	0.023	−.1477517	−.0110844
lhqdg	.121542	.0649816	1.87	0.061	−.0058196	.2489036
lhqhndA	−.0084562	.0620451	−0.14	0.892	−.1300623	.1131499
lhqnone	−.133598	.0505779	−2.64	0.008	−.2327288	−.0344672
lhqoth	−.0788673	.0755278	−1.04	0.296	−.2268991	.0691644
rural	.1165216	.0396693	2.94	0.003	.0387712	.1942721
suburb	.0519555	.0316588	1.64	0.101	−.0100945	.1140056
mothsmo	−.4478221	.0748388	−5.98	0.000	−.5945034	−.3011408
fathsmo	−.3740176	.0442889	−8.44	0.000	−.4608223	−.287213
bothsmo	−.5199595	.0506314	−10.27	0.000	−.6191953	−.4207237
smother	−.2308891	.0296456	−7.79	0.000	−.2889934	−.1727848

male	−.6421316	.0280485	−22.89	0.000	−.6971056	−.5871576
lnage	.3321544	.0764499	4.34	0.000	.1823153	.4819935
_cons	2.851352	.3175137	8.98	0.000	2.229036	3.473667
/ln_gam	−.6430285	.0157277	−40.88	0.000	−.6738543	−.6122027
gamma	.525698	.008268			.5097401	.5421554

Time to starting smoking is predicted to be shorter (‘accelerated’) for men, individuals in the lowest socioeconomic groups and those with no formal qualifications. Other smokers in the family also accelerate time to failure, meaning that the age of starting is younger for these individuals. As expected, survival time is longer for older individuals. Gamma is the shape parameter and is estimated to be positive and smaller than 1, thus indicating that the hazard first rises with survival time and then falls monotonically. It must be noted that the option time after streg produces coefficients that are equal to the coefficients estimated from the proportional hazard (PH) model divided by the ancillary parameter. In the case of the log-logistic, which has only an AFT metric, the option nohr and the option time give the same coefficients.

Next we calculate the predicted mean and median survival using the post-estimation command predict:

• predict median_start, median time
• predict mean_start, mean time
• summ mean_start median_start

Variable	Obs	Mean	Std. Dev.	Min	Max
mean_start	4646	57.00607	24.00365	19.65003	168.7198
median_start	4646	34.40472	14.48686	11.85933	101.827

We use stcurve to graph the survivor, the hazard and the cumulative hazard functions of the fitted model.

• stcurve, survival title (“LogLog Survivor-starting”) saving (logLsurvstart, replace) (note: file logLsurvstart. gph not found) (file logLsurvstart. gph saved)

• stcurve, hazard title (“LogLog Hazard-starting”) saving (logLhazstart, replace) (note: file logLhazstart. gph not found) (file logLhazstart. gph saved)

• stcurve, cumhaz title (“LogLogCumulative Hazard-starting”) saving (logLcumhazstart, replace) (note: file logLcumhazstart. gph not found) (file logLcumhazstart. gph saved)

• gr combine “logLsurvstart” “logLhazstart” “logLcumhazstart”, saving (startingP, replace) (file startingP. gph saved)

Figure 6.3 shows that the hazard of starting is higher for young individuals (lower duration) and decreases dramatically for survival times higher than 35 years, as the empirical hazard function in Figure 6.1 suggested.

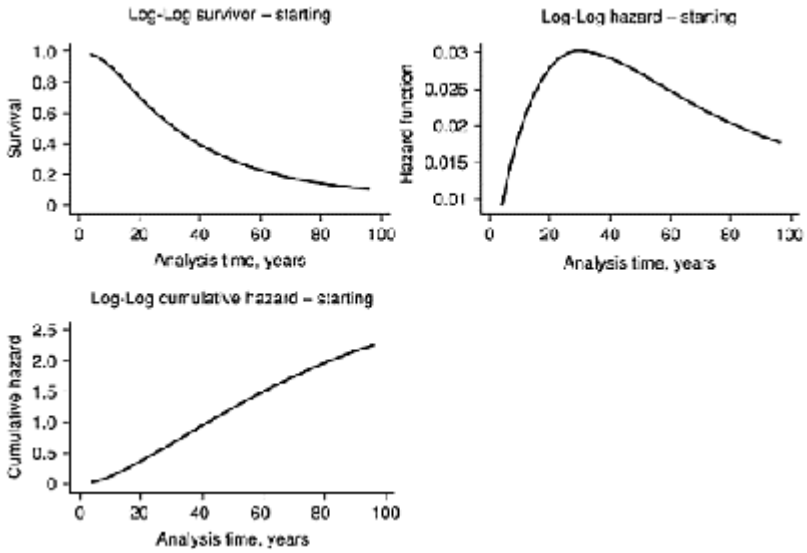

Figure 6.3 Log-logistic functions for smoking initiation.

Table 6.2 and Figure 6.3 suggest that, as it stands, the model is inadequate. The predicted age of starting is far too high when compared to the actual age of starting among smokers. Standard duration models, like the loglogistic model used here, assume that eventually everyone fails—in this case everyone would eventually start smoking. This seems to be an implausible assumption, and models based on the assumption do not do a good job of fitting the observed data. An alternative is to use a so-called split population model. This augments the standard duration analysis by adding a splitting mechanism. So, for example, a probit specification could be added to model the probability that somebody will eventually start smoking. When this splitting mechanism is added to the duration model, it does a far better job of explaining the observed data on age of starting than models that omit a splitting mechanism (see Forster and Jones 2001).

The results of Forster and Jones (2001) suggest that a simplified version of the split population model will work well with the HALS data. This uses a standard binary choice model, such as a logit or probit, for the indicator of whether an individual has started (start) and then applies the duration model only to the starters in the sample. This can be viewed as a two-part specification of the duration model.

Again we can compare different distributions for the sub-sample of 'starters' using both the Cox-Snell residuals graphical test and the information criteria. This implies specifying the option if in the stset commands as reported in the case of the exponential below:

• stset, clear
  • qui stset starting if start==1, failure(start) id(serno)
  • qui streg $xstart, d(exp) time

• estimates store exp
• predict double cs, csnell
• stset, clear
• qui stset cs if start==1, failure (start)
• qui sts gen km=s
• qui gen double H=−ln (km)
• qui line H cs cs, sort title (“Exponential”) leg (off) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“ExpCSstart1”,replace)

Replicating this for each distribution we obtain Figure 6.4 and Table 6.3.

Table 6.3 Information criteria—starters—smoking initiation

Model	Obs	11 (null)	11 (model)	df	AIC	BIC
exp	2901	−3042.621	−3019.219	15	6068.437	6158.029
weibull	2901	−1208.255	−853.0659	16	1738.132	1833.697
logN	2901	−672.9196	−443.9794	16	919.9588	1015.524
logL	2901	−498.6487	−291.3713	16	614.7426	710.3076

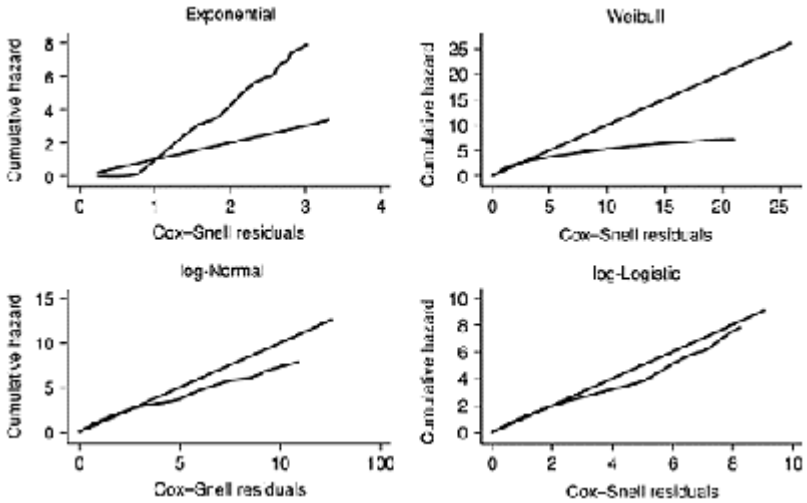

Figure 6.4 Cox-Snell residuals test—starters—smoking initiation.

The graphical test and the information criteria again favour the log-logistic distribution, which we use to estimate the final parametric model for smoking initiation:

• stset, clear

- qui stset starting if start==1, failure(start)
- stdes

The stdes command produces the following table, which shows that the mean and median survival times are 18 and 17 respectively.

failure _d: start
analysis time_t: starting

		per subject			
Category	total	mean	min	median	max
no. of subjects	2901				
no. of records	2901	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		18.04688	4	17	70
subjects with gap	0				
time on gap if gap	0				
time at risk	52354	18.04688	4	17	70
failures	2901	1	1	1	1

The estimation of the duration model on the sub-sample of starters is reported in Table 6.4:

- streg $xstart, dist (loglogistic) nolog

Table 6.4 Smoking initiation for starters—log-logistic distribution (AFT)coefficients

Log-logistic regression--accelerated failure-time form

No. of subjects=	2901	Number of obs=	2901
No. of failures=	2901		
Time at risk=	52354		
		LR chi2 (14)=	414.55
Log likelihood=	−291.37131	Prob>chi2=	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sc12	.0038026	.0121417	0.31	0.754	−.0199947	.0275999
sc45	−.0259335	.0115498	−2.25	0.025	−.0485707	−.0032962
lhqdg	.0240003	.0216664	1.11	0.268	−.0184651	.0664657
lhqhndA	.0131455	.020759	0.63	0.527	−.0275414	.0538325
lhqnone	−.0619486	.0171223	−3.62	0.000	−.0955076	−.0283895
lhqoth	−.0223088	.0253117	−0.88	0.378	−.0719188	.0273011
rural	.0145314	.0131891	1.10	0.271	−.0113187	.0403815

suburb	.0177562	.0105353	1.69	0.092	−.0028926	.0384051
mothsmo	−.0577735	.0251914	−2.29	0.022	−.1071478	−.0083992
fathsmo	−.0597182	.0156403	−3.82	0.000	−.0903727	−.0290637
bothsmo	−.08849	.0175068	−5.05	0.000	−.1228027	−.0541773
smother	−.0142472	.0097939	−1.45	0.146	−.0334429	.0049485
male	−.1707894	.0094643	−18.05	0.000	−.1893391	−.1522396
lnage	.1123013	.0263235	4.27	0.000	.0607082	.1638944
_cons	2.571063	.1089259	23.60	0.000	2.357572	2.784554
/ln_gam	−1.925379	.0158796	−121.25	0.000	−1.956502	−1.894255
gamma	.1458205	.0023156			.141352	.1504303

Time to failure is predicted to be accelerated for men, individuals from the lowest social group and with no education, and individuals whose relatives smoke. The variables lnage and suburb have the opposite effect and in fact they slow down time to starting smoking. The parameter gamma still indicates that the hazard rises before declining monotonically with survival time.

We calculate the predicted survival times:

• predict median_start, median time
• predict mean_start, mean time
• summ mean_start median start

Variable	Obs	Mean	Std. Dev.	Min	Max
mean_start	2901	17.6986	1.715094	14.29607	23.5218
median_start	2901	17.08602	1.655731	13.80125	22.70767

The predicted survival is very close to the observed value and the model fits much better for the data on the sub-sample of starters.

The following commands are used to produce the graphical analysis from the fitted model:

• stcurve, survival title (“LogLog Survivor-starting”) saving (logLsurvstart1, replace)
• stcurve, hazard title (“LogLog Hazard-starting”) saving (logLhazstart1, replace)
• stcurve, cumhaz title (“LogLog Cumulative Hazard-starting”) saving (logLcumhazstart1, replace)
• gr combine “logLsurvstart1” “logLhazstart1” “logLcumhazstart1”, saving (startingP1, replace)

Figure 6.5 shows that survival declines rapidly from ages 17–18. The hazard is predicted to rise and then fall monotonically.

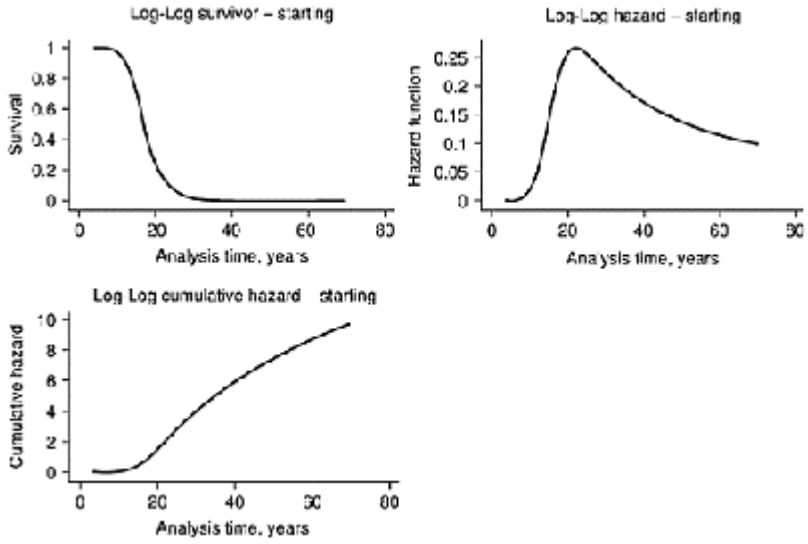

Figure 6.5 Log-logistic functions for smoking initiation.

Smoking cessation

As we did for the onset of smoking, we use some basic commands to investigate smoking duration:

• tab quit

quit	Freq.	Percent	Cum.
0	1,443	49.74	49.74
1	1,458	50.26	100.00
Total	2,901	100.00	

• su sm_years, de

failure event:	quit !=0 & quit<.	
obs. time interval:	(0, sm_years]	
exit on or before:	failure	
4646	total obs.	
1745	event time missing (sm_years>=.)	PROBABLE ERROR
2901	obs. remaining, representing	
1458	failures in single record/single failure data	
93520	total analysis time at risk, at risk from t=	0

earliest observed entry t=	0
last observed exit t=	72

• stset, clear
  • stset sm_year, failure (quit)
  • stsum

failure _d: quit
analysis time t: sm_years

	time at risk	incidence rate	no. of subject	Survival time 25%	Survival time 50%	Survival time 75%
total	93520	.0155902	2901	26	43	59

• stdes

failure _d: quit
analysis time _t: sm_years

Category	total	per subject mean	per subject min	per subject median	per subject max
no. of subjects	2901				
no. of records	2901	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		32.23716	1	32	72
subjects with gap	0				
time on gap if gap	0				
time at risk	93520	32.23716	1	32	72
failures	1458	.5025853	0	1	1

The first table shows that 50% of the 2,910 individuals who started smoking decided to stop. For this reason the use of stset needs to be conditional on the sub-sample of starters. The failure event is indicated by the dummy variable quit. The stsum command shows the cumulative distribution of total time at risk: 25% of the sample quit after 26 years, 50% after 43 years and 75% after 59 years of smoking. The stdes output shows that both mean and median time at risk is about 32 years and the longest smoking duration is 72 years.

We use a non-parametric procedure to explore survival time and the hazard function:

• sts graph if start==1, title (“Kaplan-Meier Survivor-quitting”) saving (KMsurvquit, replace)

• sts graph if start==1, hazard title (“Kaplan-Meier Hazard-quitting”) saving (KMhazardquit, replace)

• sts graph if start==1, na ylab (0 (1) 3) title (“Nelson-Aalen cumulative Hazard function-quitting”) saving (NAcumhazquit, replace)

• gr combine “KMsurvquit” “KMhazardquit” “NAcumhazquit”, saving (quittingNP, replace)

This code graphs the Kaplan-Meier survivor and hazard function and the Nelson-Aalen cumulative hazard (Figure 6.6).

The survivor function has a decreasing pattern, although it diminishes less than proportionally for the first 40 years of the analysis time. In fact,

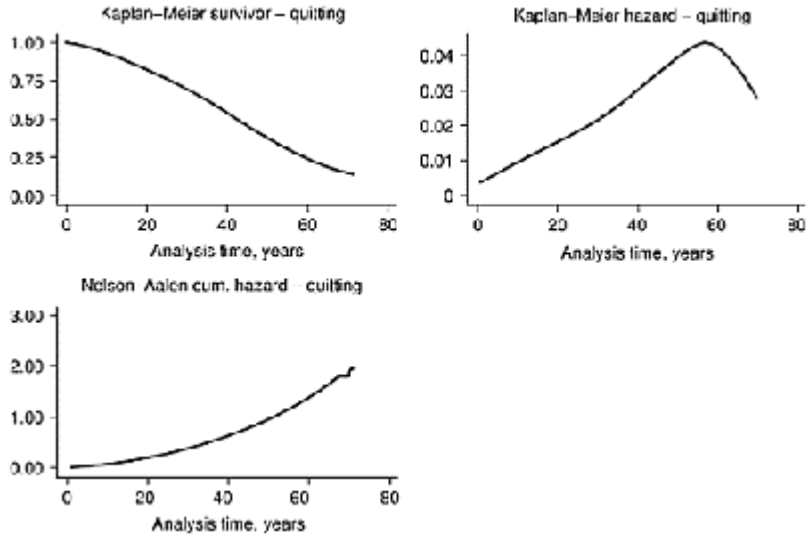

Figure 6.6 Non-parametric functions for smoking cessation.

the survival probability is still very high (between 1 and 0.50) for durations as long as 40 years and than decreases faster. The shape of the hazard function is increasing and shows a peak between 50 and 60 years of smoking, and then decreases. This is confirmed by the big jump in the cumulative hazard function at the highest survival times.

The parametric model includes a list of regressors defined as:

• global xquit “sc12 sc45 lhqdg lhqhndA lhqnone lhqoth widow sepdiv single part unemp sick retd keephse wkshf t1 rural suburb housown hou mothsmo f athsmo bothsmo smother male lnage lnyear”

The hazard of quitting is assumed to depend on a large set of socioeconomic characteristics, demographics and family smoking behaviour. Marital status is used as a proxy of social network and social support and is assumed to influence the decision to quit as well as the duration of smoking. The logarithm of year at starting is used as a control. The parametric distribution that best fits the data is chosen using the Cox-Snell residuals test and the information criteria:

/*Exponential*/

- stset, clear
- qui stset sm_year, failure (quit) id (serno)
- qui streg $xquit, d(exp) time
- estimates store exp
- predict double *cs,* csnell
- stset, clear
- qui stset cs, failure (quit)
- qui sts gen km=s
- qui gen double H=−ln (km)
- qui line H cs cs, sort title (“Exponential”) leg (off) ylabel(0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“ExpCSquit”, replace)

/*Weibull*/

- stset, clear
- qui stset sm_year, failure (quit) id (serno)
- qui streg $xquit, d(w) time
- estimates store weibull
- predict double cs2, csnell
- stset, clear
- qui stset cs2, failure (quit)
- qui sts gen km2=s
- qui gen double H2=−ln (km2)
- qui line H2 cs2 cs2, sort title (“Weibull”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“WeibCSquit”, replace)

/*Log-Normal*/

- stset, clear
- qui stset sm_year, failure (quit) id (serno)
- qui streg $xquit, d (lognormal)
- estimates store logN
- predict double cs3, csnell
- stset, clear
- qui stset cs3, failure (quit)
- qui sts gen km3=s
- qui gen double H3=−ln (km3)
- qui line H3 cs3 cs3, sort title (“logNormal”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“LogNormalCSquit”, replace)

/*Log-Logistic*/

- stset, clear
- qui stset sm_year, failure (quit) id (serno)
- qui streg $xquit, d (loglogistic)

• estimates store logL
• predict double cs4, csnell
• stset, clear
• qui stset cs4, failure (quit)
• qui sts gen km4=s
• qui gen double H4=−ln (km4)
• qui line H4 cs4 cs4, sort title (“logLogistic”) ylabel (0 (0.5) 3) legend (off) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“LogLogisticCSquit”, replace)

• drop cs km H cs2 km2 H2 cs3 km3 H3 cs4 km4 H4

• gr combine “ExpCSquit” “WeibCSquit” “LogNormalCSquit” “LogLogisticCSquit”, imargin (1 10 1 10) graphregion (margin (l=2 r=2)) saving (“CoxSnell_quit”, replace)

• estimates stats exp weibull logN logL
• drop est*

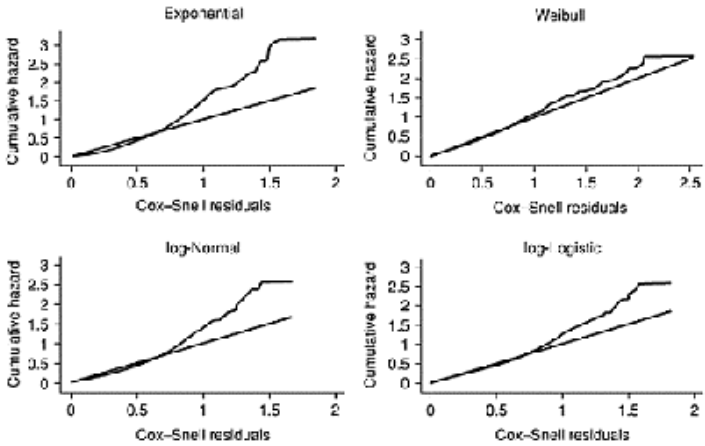

Figure 6.7 Cox-Snell residuals test–smoking cessation.

Figure 6.7 and Table 6.5 compare the exponential, Weibull, log-normal and log-logistic distributions.

Table 6.5 Information criteria–smoking cessation

Model	Obs	11(null)	11(model)	df	AIC	BIC
exponential	2901	−3012.043	−2874.715	27	5803.43	5964.696
weibull	2901	−2781.768	−2560.507	28	5177.013	5344.252
logN	2901	−2911.074	−2738.088	28	5532.176	5699.414
logL	2901	−2828.102	−2621.761	28	5299.522	5466.76

The graphical analysis shows that, except for the exponential, all the distributions fit the data quite well. The penalized log-likelihood criteria calculated by estat favour the Weibull distribution. The Weibull model is characterized by the following expressions for the hazard, the survivor and the density function:

$$h(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = \lambda_i p t_i^{p-1}$$

$$S(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = \exp(-\lambda_i t_i^{p})$$

$$f(t_i \mid \boldsymbol{x}_i;\boldsymbol{\beta}) = \lambda_i p t_i^{p-1}\exp(-\lambda_i t_i^{p})$$

Where λ_i, is a non-negative function that depends on the observed characteristics, λ_i=exp$(-px_i\beta)$; pt_i^{p-1} is the baseline hazard whose shape depends on the ancillary parameter p. The Weibull model can yield a mono tonic increasing or decreasing hazard of quitting. Regarding this, the sign of the shape parameters needs to be interpreted. If p=1 the Weibull equals the exponential, with $h(t)=\lambda$. If $p>1$ the hazard function is monotonically increasing; if $p<1$ the hazard function is monotonically decreasing. The last two cases are known as positive and negative duration dependence.

For the Weibull model the likelihood is:

$$L_i = (\lambda_i p t_i^{p-1}\exp(-\lambda_i t_i^{p}))^{quit} \cdot (\exp(-\lambda_i t_i^{p}))^{(1-quit)}$$

This expression is maximized in Stata with the command streg. We estimate a Weibull model for quitting smoking in accelerated failure time metric and save the predicted mean and median survival time (Table 6.6):

• stset, clear
  • qui stset sm_year, failure (quit)
  • streg $xquit, dist (weibull) time nolog

Table 6.6 Smoking cessation—Weibull distribution (AFT)—coefficients

Weibull regression--accelerated failure-time form

No. of subjects=	2901	Number of obs=	2901		
No. of failures=	1458				
Time at risk=	93520				
		LR chi2 (26)=	442.52		
Log likelihood=	−2560.5066	Prob>chi2=	0.0000		

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sc12	−.0867281	.0347142	−2.50	0.012	−.1547668	−.0186895
sc45	.0357743	.0349416	1.02	0.306	−.03271	.1042586

lhqdg	−.0773818	.0613357	−1.26	0.207	−.1975976	.0428341
lhqhndA	.0428267	.0603626	0.71	0.478	−.0754817	.1611352
lhqnone	.1259208	.0503522	2.50	0.012	.0272324	.2246092
lhqoth	.154064	.0748662	2.06	0.040	.007329	.300799
widow	.0452093	.0470219	0.96	0.336	−.046952	.1373706
sepdiv	.2991298	.0783928	3.82	0.000	.1454827	.4527768
single	.1302167	.063006	2.07	0.039	.0067272	.2537062
part	−.0353926	.0523524	−0.68	0.499	−.1380014	.0672163
unemp	.2480242	.0880589	2.82	0.005	.0754319	.4206164
sick	.1551065	.0758037	2.05	0.041	.0065341	.303679
retd	.0116477	.0464198	0.25	0.802	−.0793334	.1026288
keephse	.0464702	.0640859	0.73	0.468	−.0791358	.1720762
wkshft1	.1080792	.0643879	1.68	0.093	−.0181188	.2342772
rural	−.0716918	.0380537	−1.88	0.060	−.1462756	.002892
suburb	−.0566373	.031345	−1.81	0.071	−.1180723	.0047978
housown	.0445485	.053348	0.84	0.404	−.0600116	.1491086
hou	−.0386157	.0151083	−2.56	0.011	−.0682274	−.009004
mothsmo	.0103378	.0749785	0.14	0.890	−.1366173	.1572929
fathsmo	−.0137415	.0427324	−0.32	0.748	−.0974955	.0700124
bothsmo	.0198731	.0498491	0.40	0.690	−.0778293	.1175756
smother	.4133573	.0335065	12.34	0.000	.3476858	.4790288
male	−.113804	.0348075	−3.27	0.001	−.1820253	−.0455826
lnage	−.0505022	.1702438	−0.30	0.767	−.3841738	.2831694
lnyear	−.5379062	.0768245	−7.00	0.000	−.6884795	−.3873329
_cons	5.877322	.9102839	6.46	0.000	4.093199	7.661446
/ln_p	.6751269	.0238205	28.34	0.000	.6284395	.7218143
P	1.964282	.0467903			1.874683	2.058164
1/p	.5090918	.0121268			.48587	.5334236

• predict median_quit1, median time
• predict mean_quit1, mean time
• summ mean_quit1 median_quit1

Variable	Obs	Mean	Std. Dev.	Min	Max
mean_quit1	2901	44.85645	13.66769	18.69054	125.2251
median_quit1	2901	41.98409	12.79249	17.4937	117.2064

Table 6.6 reports the estimated coefficients. The coefficients for sc12 and male are negative, meaning that time to quitting smoking accelerates for individuals in the top social group and men. It also accelerates with lnage and lnyear, meaning that older individuals as well as individuals who started smoking later have a shorter survival time.

Survival in the state of smoking is also shorter for individuals from rural areas. The socioeconomic gradient in quitting is confirmed by the coefficients of the variables lhqnone lhqoth unemp sick, which suggests that for less educated individuals and persons who do not work time to quitting decelerates, hence the decision to quit is postponed. The ancillary parameter p is positive, suggesting that the hazard is monotonically increasing.

The graphical analysis allows comparison of the fitted survivor and hazard functions with the non-parametric functions, and shows that the Weibull survivor function is a good match for the Kaplan-Meier survivor function. The Weibull hazard function moves away from the empirical hazard function only for the right tail of the survival time distribution (Figure 6.8).

• stcurve, survival title ("Weibull Survivor-quitting") saving (wsurvquit1, replace) (file wsurvquit1. gph saved)

• stcurve, hazard title ("Weibull Hazard-quitting") saving (wHquit1, replace) (file wHquit1. gph saved)

• stcurve, cumhaz title ("Weibull Cumulative Hazard-quitting") saving (wcumHquit1, replace) (file wcumHquit1. gph saved)

• gr combine "wsurvquit1" "wHquit1" "wcumHquit1", saving ("quittingP1", replace) (file quittingP. gph saved)

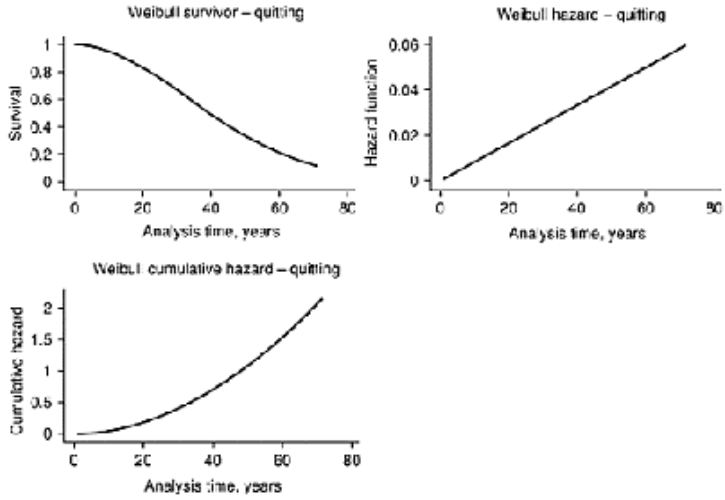

Figure 6.8 Weibull estimated functions for smoking cessation.

Lifespan

To study the hazard of dying in the HALS sample we use the same approach used so far with smoking duration. Deaths account for about 43% of the sample of those aged over 40 at HALS. We first calculate some statistics for the survival time variable lifespan:

• su lifespan, de

Survivor time: censoring at April 2005

	Percentiles	Smallest		
1%	55.1	41.5		
5%	60.55852	43.9		
10%	61.82615	44.4	Obs	4646
25%	66.2	45.4	Sum of Wgt.	4646
50%	73.3		Mean	73.99948
		Largest	Std. Dev.	9.62414
75%	81.2	101.6		
90%	86.8	106.1602	Variance	92.62407
95%	90.1	106.2834	Skewness	.2362737
99%	96.47365	110.976	Kurtosis	2.523947

This shows that 1% of the sample survived to age 55, and half of the sample up to 73. The mean survival time is about 74.

The pnorm command produces a graph of a standardized normal probability (normal probability plot) and suggests that lifespan is probably distributed as a Gaussian random variable (Figure 6.9):

• pnorm lifespan

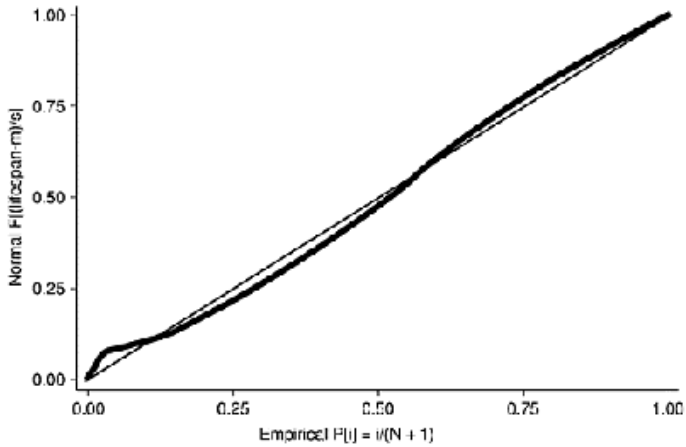

Figure 6.9 Normal probability plot for lifespan.

We use the stset command to explore the survival time for lifespan. The event of failure is death and age is used as the variable that indicates left-truncation.

- stset, clear
- stset lifespan, failure (death) enter (age)

```
        failure event: death !=0 & death<.
obs. time interval: (0, lifespan]
 enter on or after: time age
exit on or before: failure
------------------------------------------------------------------------------
      4646  total obs.
         0  exclusions
------------------------------------------------------------------------------
      4646  obs. remaining, representing
      1987  failures in single record/single failure data
   74121.4  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =        40
                                 last observed exit t =   110.976
```

Here the key concept is that individuals who died before 1984 (the observation time in HALS) are not surveyed but are excluded from the sample. The idea of exclusion must not be confounded with the problem of missing observations, but simply refers to the fact that the HALS sample is the result of a selection process, conditional on the event of death having not occurred prior to the survey time. We could think of the HALS sample as made up of individuals who have a relative lower hazard of dying, since individuals with a higher hazard left the initial state before any information had been collected about them. Hence, the remaining sample has a lower hazard relative to the truncated part of the population. The option enter allows us to account for left truncation and should not be confounded with options origin and time0. The Stata 9 manual [ST] briefly illustrates the main differences between options in stset.

- stsum

```
        failure _d: death
analysis time _t: lifespan
 enter on or after: time age
```

	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
total	74121.40066	.0268074	4646	71.6	80.2	87.3

We learn that 25% of the sample survive up to age 72 and 50% up to age 80.

- stdes

failure _d: death
analysis time _t: lifespan
enter on or after: time age

Category	total	mean	per subject: min	median	max
no. of subjects	4646				
no. of records	4646	1	1	1	1
(first) entry time		58.04567	40	57.2	96.8
(final) exit time		73.99948	41.5	73.3	110.976
subjects with gap	0				
time on gap if gap	0				
time at risk	74121.401	15.95381	.0999985	19.79925	20.53997
failures	1987	.4276797	0	0	1

This shows that a different entry time than the time at which individuals start to be at risk of failure has been specified. Notice the difference between the command used here and those used for the smoking data in the line (first) entry time, because here summary statistics are calculated also for survival time at entry. On average, survival time is 58 at entry and 74 at exit. Statistics for survival time at entry can be verified with the following command:

• sum age

Variable	Obs	Mean	Std. Dev.	Min	Max
age	4646	58.04567	11.75395	40	96.8

The usual non-parametric procedures are used:

. sts graph, title (“Kaplan-Meier Survivor-lifespan”) saving (KMsurvls, replace)
. sts graph, hazard title (“Kaplan-Meier Hazard-lifespan”) saving (KMhazardls, replace)
. sts graph, na title (“Nelson-Aalen cumulative Hazard function-lifespan”) saving (NAcumhazls, replace)
. gr combine “KMsurvls” “KMhazardls” “NAcumhazls”, saving (lifespanNP, replace)

The combined graph is shown in Figure 6.10.

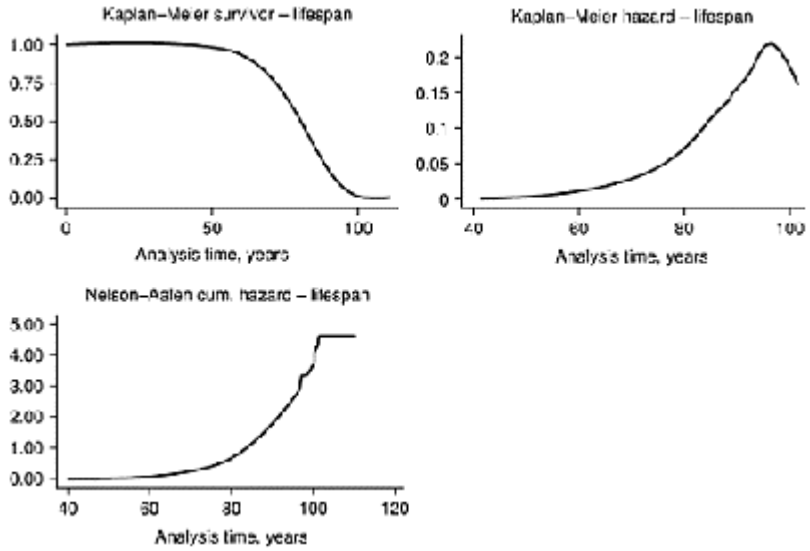

Figure 6.10 Non-parametric functions for lifespan.

Figure 6.10 shows that the survivor function has a decreasing pattern and the hazard function is increasing, but with different slopes: it is steeper for durations longer than 80 years and shows a decreasing pattern after the peak, around age 95 where the cumulative hazard function appears to be flat.

For the parametric analysis we define a global for the regressors:

• replace quit=0 if quit==.

• global xls "sc12 sc45 lhqdg lhqhndA lhqnone lhqoth part unemp sick retd keephse wkshf t1 rural suburb male lnage start quit"

The hazard of dying is assumed to be a function of socioeconomic characteristics and demographics. Smoking behaviour is considered an important determinant of the hazard so the variables start and quit are included as regressors. In order to estimate the model for lif espan on the full sample, missing values for quit are set equal to zero. We do not attempt to deal with potential endogeneity of these variables, so the estimates should be treated with caution if they are to be interpreted as causal effects of smoking on lifespan.

We calculate the Cox-Snell residual test and the information criteria to compare five alternative distributions (Figure 6.11 and Table 6.7). We include the Gompertz distribution because the law of human mortality was first described by Benjamin Gompertz in 1825, and the Gompertz mortality model has been used in biology and medical modelling:

/*Exponential*/

• stset, clear

- qui stset lifespan, failure (death) enter (age)
- qui streg $xls d(exp) time
- estimates store exp
- predict double cs, csnell
- stset, clear
- qui stset cs, failure (death)
- qui sts gen km=s
- qui gen double H=−ln (km)
- qui line H cs cs, sort title (“Exponential”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“ExpCSls”, replace)

/*Weibull*/

- stset, clear
- qui stset lifespan, failure (death) enter (age)
- qui streg $xls, d(w) time
- estimates store weibull
- predict double cs2, csnell
- stset, clear
- qui stset cs2, failure (death)
- qui sts gen km2=s
- qui gen double H2=−ln (km2)
- qui line H2 cs2 cs2, sort title (“Weibull”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“WeibCSls”, replace)

/*Log-Normal*/

- stset, clear
- qui stset lifespan, failure (death) enter (age)
- qui streg $xls, d(lognormal)
- estimates store logN
- predict double cs3, csnell
- stset, clear
- qui stset cs3, failure (death)
- qui sts gen km3=s
- qui gen double H3=−ln (km3)
- qui line H3 cs3 cs3, sort title (“logNormal”) leg (off) ylabel (0 (0.5) 3) ytitle (“Cumulative Hazard”, size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving (“LogNormalCSls”, replace)

/*Log-Logistic*/

- stset, clear
- qui stset lifespan, failure (death) enter (age)
- qui streg $xls, d (loglogistic)
- estimates store logL
- predict double cs4, csnell

- stset, clear
- qui stset cs4, failure (death)
- qui sts gen km4=s
- qui gen double H4=−ln (km4)

/*Gompertz*/

- stset, clear
- qui stset lifespan, failure (death) enter (age)
- qui streg $xls, d(gompertz)
- estimates store gomp
- predict double cs5, csnell
- stset, clear
- qui stset cs5, failure (death)
- qui sts gen km5=s
- qui gen double H5=−ln (km5)
- qui line H5 cs5 cs5, sort title ("Gompertz") ylabel (0 (0.5) 3) legend (off) ytitle ("Cumulative Hazard", size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving ("GompertzCSls", replace)

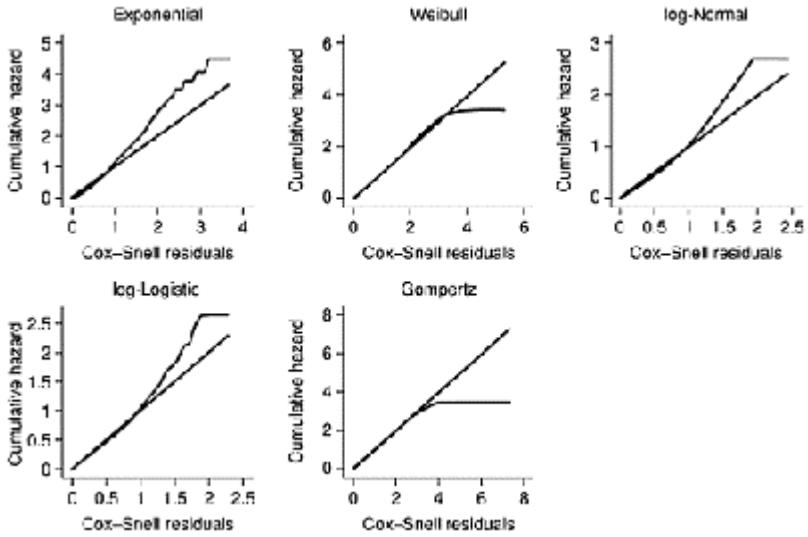

Figure 6.11 Cox-Snell residuals test—lifespan.

- qui line H4 cs4 cs4, sort title ("logLogistic") ylabel (0 (0.5) 3) legend (off) ytitle ("Cumulative Hazard", size (small) margin (medsmall)) xtitle (Cox-Snell Residuals, size (small) margin (medsmall)) saving ("LogLogisticCSls", replace)

- drop cs km H cs2 km2 H2 cs3 km3 H3 cs4 km4 H4 cs5 km5 H5

• gr combine “ExpCSls” “WeibCSls” “LogNormalCSls” “LogLogisticCSls” “GompertzCSls”, imargin(1 10 1 10) graphregion (margin (l=2 r=2)) saving (“CoxSnell_1s”, replace)

• estimates stats exp weibull logN logL gomp

• drop_est*

Table 6.7 Information criteria—lifespan

Model	Obs	ll (null)	ll (model)	df	AIC	BIG
exp	4646	−577.0944	487.887	19	−937.7741	−815.3426
weibull	4646	521.9582	710.9347	20	−1381.869	−1252.994
logN	4646	440.1666	620.3466	20	−1200.693	−1071.818
logL	4646	461.7225	643.5775	20	−1247.155	−1118.28
gomp	4646	510.0126	709.284	20	−1378.568	−1249.693

The estimation of a Weibull model for lifespan is not successful because the maximization process does not converge to a global maximum: Stata reports the error message convergence not achieved. So the tests based on the Weibull estimated coefficient and residuals are not reliable. Both the Cox-Snell residuals test and the information criteria favour the Gompertz distribution for lifespan. The Gompertz model is parameterized as follows:

$$h(t_i \mid \boldsymbol{x}_i;\beta) = \lambda_i \exp(\gamma t_i)$$

$$S(t_i \mid \boldsymbol{x}_i;\beta) = \exp\left\{-\frac{\lambda_i}{\gamma}[\exp(\gamma t_i) - 1]\right\}$$

$$f(t_i \mid \boldsymbol{x}_i;\beta) = \lambda_i \exp(\gamma t_i) \cdot \exp\left\{-\frac{\lambda_i}{\gamma}[\exp(\gamma t_i) - 1]\right\}$$

The likelihood for a left-truncated duration model is given by:

$$L_i = (h(t_i \mid \boldsymbol{x}_i;\beta))^{d_i} \cdot \left(\frac{S(t_i \mid x_i;\beta)}{S(\tau_i \mid x_i;\beta)}\right)$$

Where individuals with a complete spell (death==1) contribute to the likelihood with their hazard function and individuals with right-censored spells (death==0) contribute with their survivor function conditional on survival up to the interview date, τ_j.

$$L_i = (\lambda_i \exp(\gamma t_i))^{d_i} \cdot \left(\frac{\exp\left\{-\frac{\lambda_i}{\gamma}[\exp(\gamma t_i) - 1]\right\}}{\exp\left\{-\frac{\lambda_i}{\gamma}[\exp(\gamma \tau_i) - 1]\right\}} \right)$$

The Gompertz model is parameterized as a proportional hazard model or log-relative hazard form:

$h(t_i|x_i;\beta)=h_0(t)\cdot\exp(x_i\beta)$

where the baseline hazard is $h_0(t)=\exp(\gamma t_i)$ and $\exp(x_i\beta)=\lambda_i$ scales the baseline hazard multiplicatively by the same amount at each instant t. The ancillary parameter γ is estimated by Stata. If $\gamma>0$ the hazard function increases with time, if $\gamma<0$ the hazard function decreases with time. The exponential hazard function is a special case of the Gompertz hazard when $\gamma=0$.

We use streg to maximize the log-likelihood function with options nohr and hr to get

$$\exp(\beta_k) = \frac{h(t^{\%}, x_i)}{h(t^{\%}, x_j)};$$

both the coefficients β_k and the hazard ratios

- stset, clear
  - stset lifespan, failure (death) enter (age)
  - streg $xls, d(gompertz) nohr nolog

See Tables 6.8 and 6.9.

Table 6.8 Lifespan—Gompertz model—coefficients

Gompertz regression--log relative-hazard form

No. of subjects=	4646		Number of obs=		4646
No. of failures=	1987				
Time at risk=	74121.40066				
			LRchi2 (18)=		398.54
Log likelihood=	709.28401		Prob>chi2=		0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sc12	−.1530234	.0625764	−2.45	0.014	−.2756709	−.030376
sc45	.1032688	.0540068	1.91	0.056	−.0025826	.2091202
lhqdg	−.0569884	.1220213	−0.47	0.640	−.2961457	.1821689

lhqhndA	−.1958631	.1234852	−1.59	0.113	−.4378897	.0461634
lhqnone	.1045263	.0961881	1.09	0.277	−.0839989	.2930516
lhqoth	.1157708	.1332891	0.87	0.385	−.145471	.3770126
part	.0273555	.1041368	0.26	0.793	−.1767489	.23146
unemp	.4617398	.1428298	3.23	0.001	.1817986	.7416811
sick	.7548937	.1163245	6.49	0.000	.5269019	.9828854
retd	.0303732	.0859395	0.35	0.724	−.1380651	.1988115
keephse	.2416043	.1226376	1.97	0.049	.001239	.4819696
wkshft1	−.1831301	.147969	−1.24	0.216	−.4731439	.1068838
rural	−.0867819	.064179	−1.35	0.176	−.2125705	.0390066
suburb	−.0672467	.0506829	−1.33	0.185	−.1665834	.03209
male	.4090663	.0518999	7.88	0.000	.3073444	.5107882
lnage	.971717	.2983871	3.26	0.001	.3868891	1.556545
start	.6437472	.0594012	10.84	0.000	.527323	.7601713
quit	−.3777534	.0563624	−6.70	0.000	−.4882218	−.2672851
_cons	−14.30188	1.011103	−14.14	0.000	−16.28361	−12.32016
gamma	.0869282	.0041305	21.05	0.000	.0788326	.0950238

• streg, hr

Table 6.9 Lifespan—Gompertz model—hazard ratios

Gompertz regression--log relative-hazard form

No. of subjects=	4646	Number of obs=	4646
No. of failures=	1987		
Time at risk=	74121.40066		
		LR chi2 (18)=	398.54
Log likelihood=	709.28401	Prob>chi2=	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sc12	8581096	.0536974	−2.45	0.014	.7590627	9700807
sc45	1.108789	.0598822	1.91	0.056	.9974207	1.232593
lhqdg	.944605	.1152619	−0.47	0.640	.743679	1.199817
lhqhndA	8221247	.1015202	−1.59	0.113	.645397	1.047246
lhqnone	1.110185	.1067866	1.09	0.277	.9194322	1.340512
lhqoth	1.122739	.1496488	0.87	0.385	.864615	1.457923
part	1.027733	.1070249	0.26	0.793	.8379902	1.260439
unemp	1.586832	.226647	3.23	0.001	1.199373	2.099462
sick	2.127385	.247467	6.49	0.000	1.693677	2.672155
retd	1.030839	.0885898	0.35	0.724	.871042	1.219952
keephse	1.27329	.1561533	1.97	0.049	1.00124	1.619261

wkshft1	8326599	.1232078	−1.24	0.216	.6230404	1.112805
rural	.916877	.0588443	−1.35	0.176	.8085033	1.039777
suburb	.9349645	.0473867	−1.33	0.185	.8465522	1.03261
male	1.505412	.0781307	7.88	0.000	1.359809	1.666604
lnage	2.642478	.7884812	3.26	0.001	1.472393	4.742408
start	1.903601	.1130761	10.84	0.000	1.69439	2.138643
quit	.6853995	.0386308	−6.70	0.000	.6137168	.7654548
gamma	0869282	.0041305	21.05	0.000	.0788326	0950238

• predict median_ls, median time
  • summ median_ls
  • drop median_ls

Variable	Obs	Mean	Std. Dev.	Min	Max
median_ls	4646	81.23831	5.72825	62.4798	96.35618

The coefficients of sc45, unemp, sick and keephse are positive, meaning that for individuals in these groups the hazard of dying is higher. For example the hazard of dying for sc45 is about 11% higher than in the other social classes. Also the coefficient of sc12 is statistically significant and, as expected, it has negative sign. The result for male indicates that, for men, the hazard of dying is 51% higher than for women.

Smoking decisions affect the risk of mortality. The choice variable start is statistically significant with a positive coefficient, meaning that the decision to start smoking increases the hazard of dying. The variable quit is statistically significant with a negative coefficient, meaning that the decision to quit slows down the hazard of dying: at each time the hazard rate of quitters is 69% of the hazard rate of those who do not quit. For a more meaningful interpretation of the coefficients we can divide the sample of respondents into current smokers, ex-smokers and never-smokers, depending on the value of start and quit. For current smokers (start==1 and quit==0) the coefficient of interest is 0.644. For ex-smokers (start==1 and quit==1) we sum the coefficients of start and quit and find that 0.266 is the effect on life span. Both current and ex-smokers have shorter lifespan than never-smokers, but the effect on the hazard of dying is bigger for those individuals who have not yet quit.

For the variable ln_age the model in nohr gives the elasticity of the hazard with respect to age, which is around 97%. The ancillary parameter gamma is positive, thus suggesting that the hazard function is increasing with time. Predicted mean survival time cannot be calculated for the Gompertz model, because there is no closed-form expression of it, but the predicted median time is about 81 years.

The survivor, hazard and cumulative hazard functions are produced by the commands below:

• stcurve, survival title (“Gompertz Survivor-lifespan”) saving (gsurvls, replace)
  • stcurve, hazard title (“Gompertz Hazard-lifespan”) saving (gHls, replace)
  • stcurve, cumh title (“Gompertz Cumulative Hazardlifespan”) saving (gcumHls, replace)

• gr combine “gsurvls” “gHls” “gcumHls”, saving (lifespanPG, replace)

The Gompertz model (Figure 6.12) produces estimated functions that mimic very well the empirical functions reported in Figure 6.10. The shape of the hazard function increases with time.

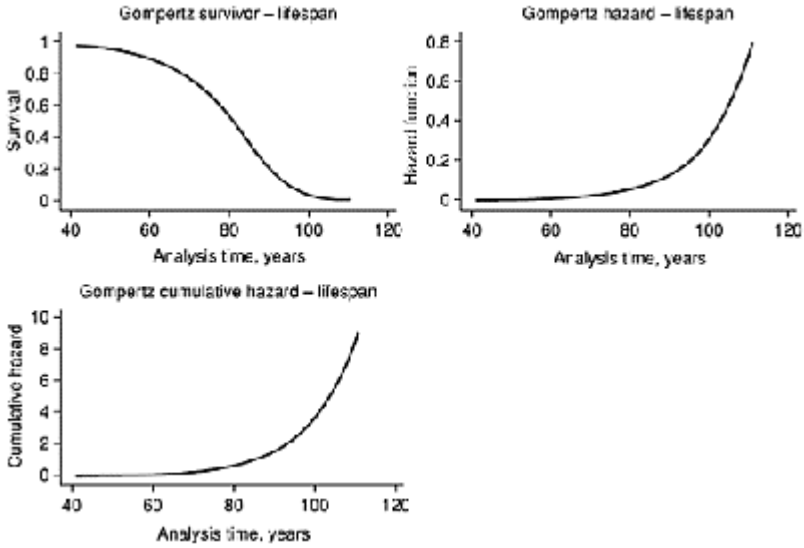

Figure 6.12 Gompertz estimated functions for lifespan.

7
Health and retirement

7.1 INTRODUCTION

This chapter illustrates the use of discrete-time duration models by analysing the impact of health on the decision to retire. Health is undoubtedly an important factor in the decision to retire. A recent survey for the Department for Work and Pensions (Humphrey *et al.* 2003) explored the factors affecting labour market participation among 2,800 people aged between 50 and 69.50% of the sample stated that they were not seeking work owing to ill health, and 20% reported that they had been forced to retire or leave a job because of ill health. However, the relationship between health and retirement is complex. It is difficult to estimate a true causal effect because health and work are jointly determined and there are problems finding an appropriate measure of health for use in this context. In order to usefully investigate the relationship it is necessary to use longitudinal data to enable us to track individuals from work into retirement, thus providing an appropriate counterfactual.

Attention is paid to the problem of potential measurement error in using self-reported measures of health status and to the question of whether a change in labour market status is best identified by a 'shock' to an individual's health or by a levels effect (for example, a slow deterioration in health status). A further issue of interest is that the majority of people in this age group live as a couple, and decisions on when to retire are often taken at the household level. Hence we also consider the effect of spousal health and labour market status on an individual's decision to retire.

7.2 PREPARING AND SUMMARIZING THE DATA

We use data from the first 12 waves (1991–2002) of the British House-hold Panel Survey (BHPS). The main variables used in the analysis are reported below.

Retirement and labour market status

The definition of retirement used here is a self-reported classification based on the answer to the question on job status (jbstat) in the BHPS. Individuals are asked to classify their status as one of the following: self-employed, employed, unemployed, retired, on maternity leave, caring for the family, in full-time education, long-term sick or disabled, or on a government training scheme. The following commands were used to recode the job status variable into a series of dummy variables, including one representing individuals who have reported themselves as retired (retired).

• /* Job status */
  • recode jbstat −9 −8 −7 −1 =. /* remove missing data */ (3106 changes made)
  • tab jbstat, gen (jobdm)

jbstat	Freq.	Percent	Cum.
self-employed	1341	12.66	12.66
employed	4766	45.01	57.67
unemployed	402	3.80	61.47
retired	3487	32.93	94.40
maternity leave	1	0.01	94.41
family care	210	1.98	96.39
ft studt, school	6	0.06	96.45
It sick, disabld	347	3.28	99.73
gvt trng scheme	3	0.03	99.75
other	26	0.25	100.00
Total	10589	100.00	

• ren jobdm1 selfemp
  • ren jobdm2 emp
  • ren jobdm3 unemp
  • ren jobdm4 retired
  • ren jobdm5 matleave
  • ren jobdm6 famcare
  • ren jobdm7 student
  • ren jobdm8 ltsick
  • ren jobdm9 govtrain
  • ren jobdm10 jobothr

For the duration analysis reported in Section 7.4 we assume that retirement is an absorbing (permanent) state. Accordingly, we follow individuals from work to the time when they first report retirement. Any subsequent transitions back to work are ignored.

Health variables

The BHPS includes a number of health and health-related variables. Of particular interest is the measure of general self-assessed health (SAH) status and an alternative measure of health that refers to limitations in daily activities.

The simple five-point SAH variable (hlstat) available in the BHPS is a subjective measure of general health based on answers to the question: ‘Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?’.

A continuity problem arises with using this variable because in wave 9 (only) there was a change in the question together with a modification to the available response categories. The wave-9 question (hlf s1) asks respondents about their ‘general state of

health' (without the age benchmark used in the original version) on the scale: excellent, very good, good, fair, poor. Both versions of the variable are coded such that 1 represents the best possible health and 5 represents worst health. We recode the variables to be increasing in good health:

```
• /* Self-assessed health */
    • recode hlstat −9 −8 −7 −6 −2 −1=. /* remove missing data */
    • recode hlstat 1=52=44=25=1/* recode variable */

• /* Change labelling of variable */
    • label define health 1 "very poor" 2 "poor" 3 "fair" 4 "good" 5 "excellent"
    • label values hlstat health
    • table hlstat

• /* General state of health */
    • recode hlsf1 −9 −8 −7 −1=.
    • recode hlsf 1 1=52=44=25=1

• /* Change labelling of variable */
    • label define health2 1 "poor" 2 "fair" 3 "good" 4 "very good" 5 "excellent"
    • label values hlsf1 health2
```

In order to maximize the span of data available to us and achieve consistency over all 12 waves we follow the method of Hernández-Quevedo *et al*. (2004) and collapse SAH into the following four-category scale where 1 represents *very poor or poor health,* 2 *fair health,* 3 *good or very good health* and 4 *excellent health* (hlstatc4). In this way both the original SAH question asked of respondents in waves 1 to 8 and 10 to 12 and the wave 9 version of the SAH question can be used.

```
    • /* Recede general health into a 4 category variable */
    • gen hlstatc4=hlstat
    • replace hlstatc4=hlsf 1 if wavenum==9
    • recode hlstatc4 2=1 if wavenum~=9
    • recode hlstatc4 3=2 if wavenum~=9
    • recode hlstatc4 4=3 if wavenum~=9
    • recode hlstatc4 5=4 if wavenum~=9
    • recode hlstatc4 4=3 if wavenum==9
    • recode hlstatc4 5=4 if wavenum==9

    • label define healthc4 1 "vpoor or poor" 2 "fair" 3 "good or vgood" 4 "excellent"
    • label values hlstatc4 healthc4
```

We can create dummy variables representing each health state as follows:

```
    • tab hlstatc4, gen (hlc4dm)
    • ren hlc4dml sah4vpp
```

- ren hlc4dm2 sah4fair
- ren hlc4dm3 sah4gvg
- ren hlc4dm4 sah4ex

Our alternative health measure is self-reported functional limitations, based on the question 'Does your health in any way limit your daily activities compared to most people of your age?'. This is arguably more objective than the general SAH question, and more directly related to ability to work, and accordingly is a useful alternative to the self-assessed health variable. The question was not asked in wave 9 and—given that health limitations are likely to consist of chronic problems—we assume that wave 8 values hold for wave 9. We create a variable coded 1 if an individual reports a health limitation and 0 otherwise (hll It yes).

- /* Health limits daily activities */
  - recode hllt −9/−1−.

- sort pid wavenum
  - replace hllt=hllt [_n−1] if wavenum==9
  - tab hllt, gen (hlltdm)
  - ren hlltdm1 hlltyes

Finally, we make use of questions on specific health problems. These are used to construct a latent health stock (see Section 7.3). Individuals are asked whether or not they have any of a list of specific health problems from the following: arms, legs or hands (hlparms), sight (hlpsee), hearing (hlphear), skin conditions or allergies (hlpskin), chest/breathing (hlpchest), heart/blood pressure (hlpheart), stomach or digestion (hlpstom), diabetes (hlpdiab), anxiety or depression (hlpanx), alcohol or drugs (hlpalch), epilepsy or migraine (hlpanx), or other (hlpothr). We create a binary dummy variable for the presence or not of each specific problem.

Spousal/partner variables

We model the impact of health on the timing of retirement separately for men and women. For both we include a variable representing the health status of the individual's spouse or partner (should they have one—shlltyes, slatsah). This allows us to investigate the interaction between spousal or partner's health and an individual's decision to retire. We also include a variable representing whether a spouse or partner is employed (lspjb). To reduce concerns over endogeneity bias this variable is lagged one period (the variable label has the prefix 1, to denote a lag).

Income and wealth

The main income variable used is the log of household income across all waves in which an individual is observed. Household income consists of labour and non-labour income (fihhyr), adjusted using the Retail Price Index and equivalized by the McClement's scale (fieqfca) to adjust for household size and composition. In the models reported here we

adapt this to represent the mean across all waves prior to retirement. This is to reduce concerns over endogeneity, as income is expected to reduce significantly at retirement (m2lnhinc) and is computed as follows:

- gen fihhyr2=fihhyr
- replace fihhyr2=fihhyr* (133.5/138.5) if wavenum==2
- replace fihhyr2=fihhyr* (133.5/140.7) if wavenum==3
- replace fihhyr2=fihhyr* (133.5/144.1) if wavenum==4
- replace fihhyr2=fihhyr* (133.5/149.1) if wavenum==5
- replace fihhyr2=fihhyr* (133.5/152.7) if wavenum==6
- replace fihhyr2=fihhyr* (133.5/157.5) if wavenum==7
- replace fihhyr2 =fihhyr* (133.5/162.9) if wavenum==8
- replace fihhyr2=fihhyr* (133.5/165.4) if wavenum==9
- replace fihhyr2=fihhyr* (133.5/170.3) if wavenum==10
- replace fihhyr2=fihhyr* (133.5/173.3) if wavenum==11
- replace fihhyr2=fihhyr* (133.5/176.2) if wavenum==12
- sort pid wavenum
- quietly by pid: gen increeq=fihhyr2/fieqf ca
- quietly by pid: gen lninc=ln(increeq)
- by pid: egen m2lnhinc=mean (lninc) if retired==0
- replace m2lnhinc=m2lnhinc [_n−1] if retired==1

We also have information on pension entitlement, which distinguishes between people who have no occupational or private pension, an occupational pension, or a private pension. From these we construct a variable representing whether an individual has ever, over the course of BHPS observations, made contributions to a private pension plan (everppenr) and whether an individual has been a member of an occupational pension plan (everemppr). Data on housing tenure are also available, which distinguish between people who own their home outright (HseOwn), own with a mortgage (HseMort), or live in privately rented (HseRent) or local authority rented housing (HseAuthAss).

Other socio-demographic variables

Other variables we are interested in using include age, sex, marital status (marcoup), educational attainment (deghdeg, hndalev, ocse), and regional dummies (northw—wales). We also include variables that indicate the employment sector of the individual in the first wave of observation (privcomp0, civlocgov0, jbsecto0). The latter variables carry the postfix 0 to indicate that they represent initial values. These variables have been constructed from their respective source variables in the BHPS. Variable names and definitions are summarized in Table 7.1.

Table 7.1 Variable names and definitions

Variable	*Description*
retired	Binary dependent variable,=1 if respondent states they are retired, 0 otherwise
hll It yes	Self-assessed health limitations: 1 if health limits daily activities, 0 otherwise

sah	Self-assessed health; 1: very poor or poor, 2: fair, 3: good or very good, 4: excellent
sah4ex	Self-assessed health: 1 if excellent, 0 otherwise
sah4gvg	Self-assessed health: 1 if good or very good, 0 otherwise
sah4fair	Self-assessed health: 1 if fair, 0 otherwise
sah4vpp	Self-assessed health: 1 if poor or very poor, 0 otherwise (baseline category)
m2lnhinc	Individual-specific mean of log equivalized real household labour and non-labour income
HseOwn	1 if house owned outright, 0 otherwise (baseline category)
HseMort	1 if house has outstanding mortgage, 0 otherwise
Variable	*Description*
HseRent	1 if house is rented, 0 otherwise
HseAuthAss	1 if house is owned by housing authority/association, 0 otherwise
marcoup	1 if married or living as a couple, 0 otherwise
deghdeg	1 if highest educational attainment is degree or higher degree, 0 otherwise
hndalev	1 if highest educational attainment is HND or A level, 0 otherwise
ocse	1 if highest educational attainment is O level or CSE, 0 otherwise
noqual	1 if no qualifications, 0 otherwise (baseline category)
everppenr	1 if respondent has made contributions to a private pension plan during observation period, 0 otherwise
everemppr	1 if respondent has been a member of an occupational pension plan during observation period, 0 otherwise
privcomp0	1 if respondent's sector of employment is within the private sector, 0 otherwise
civlocgov0	1 if respondent's sector of employment is within civic or local government, 0 otherwise
jbsecto0	1 if respondent's sector of employment is other to above, 0 otherwise
selfemp	1 if respondent is self-employed, 0 otherwise (baseline category)
job	1 if respondent's spouse/partner has a job, 0 otherwise
age5054	1 if respondent is aged 50 to 54 (inclusive), 0 otherwise
age5559	1 if respondent is aged 55 to 59 (inclusive), 0 otherwise
age6064	1 if respondent is aged 60 to 64 (inclusive), 0 otherwise
age6569	1 if respondent is aged 65 to 69 (inclusive), 0 otherwise
NorthW	1 if respondent resides in North West, Merseyside or Greater Manchester, 0 otherwise
NorthE	1 if respondent resides in North, South Yorkshire, West Yorkshire, North Yorkshire, Humberside or Tyne & Wear, 0 otherwise
SouthE	1 if respondent resides in South East or East Anglia, 0 otherwise (baseline category)
SouthW	1 if respondent resides in South West, 0 otherwise
London	1 if respondent resides in Inner or Outer London, 0 otherwise
Midland	1 if respondent resides in East or West Midlands or West Midc, 0 otherwise
Scot	1 if respodent resides in Scotland, 0 otherwise
Wales	1 if respondents resides in Wales, 0 otherwise

hlthprb Self-reported health problems: 1 if problem reported, 0 otherwise. There are also individual dummies for problems with: arms, legs or hands (arms), sight (see), hearing (hear), skin conditions or allergies (skin) chest/breathing (chest), heart/blood pressure (heart), stomach or digestion (stomach), diabetes (diabetes), anxiety or depression (anxiety), alcohol or drugs (alcohol), epilepsy (epilepsy), migraine (migraine) or Other (other).

Stock sample

Interest focuses on the role of health in determining the timing of the decision to retire. As such we wish to observe individuals who at the beginning of the BHPS survey can be considered to be at risk of retirement. Jenkins (1995) defines such a sample as a stock sample. For our purposes the stock sample consists of those individuals who were original BHPS sample members aged 50 or over *and* had provided a full interview (BHPS variable: ivfio=1) *and* were in work (defined here as employed or self-employed) in the *first wave* of the survey. This sample consists of n=1135 individuals, 494 women and 641 men. 661 individuals are present for all 12 waves, but others are lost because of sample attrition and death. Our models of retirement are estimated on complete sequences of observations such that should an individual leave the panel but then return at a later date, we only make use of information up to the wave of first exit. The Stata code for the stock sample selection mechanism is as follows:

```
• /* 1. Select if provided full interview in wave 1 */
    • drop if (ivfio~=1 & wavenum==1)

• /* 2. Select if aged 50 or over in wave 1 */
    • drop if (age<=49 & wavenum==1)

• /* 3. Select iff employed or self-employed in wave 1 */
    • drop if ( (jbstat<1|jbstat>2) & wavenum==1)

• /* 4. Select only individuals interviewed at wave 1 */

• sort pid wavenum
    • by pid: egen minwave=min (wavenum)
    • drop if minwave~=1

• /* 5. Select complete sequences of responses−i.e. stop at */
    • /* first unit non-response */
    • drop if (ivfio<1|ivfio>3)
    • gen const=1
    • by pid: gen sumcon=sum (const)
    • by pid: gen diff=wavenum−sumcon
    • drop if diff~=0
```

For illustrative purposes, this chapter only considers the retirement behaviour of men and accordingly we remove women from our sample of interest:

• drop if male~=1

Labour market transitions

The stock sample consists of individuals in work at the first wave of the BHPS. At subsequent waves, transitions to other labour market states, including retirement, may be made. We can summarize information on labour market transitions in each wave by using the tabulate command. For example, at wave 2:

• table jbstat if wavenum==2

jbstat	Freq.
self-employed	138
employed	349
unemployed	19
retired	50
It sick, disabld	11
other	2

An efficient way to summarize transitions across all waves is provided by:

• forvalues j=1(1)12{
  table jbstat if wavenum==‘j’
  }

Table 7.2 summarizes the transitions across the 12 waves and further includes individuals lost to attrition and death (where this is known in the BHPS). Our sample consists of 641 men reporting employment or self-employment status in wave 1 and this figure gradually decreases to 100 by the 12th wave. Fifty individuals classify themselves as retired in wave 2 and this increases to 241 in wave 12, representing a near five-fold increase. Some caution is required, however, in interpreting the retirement figures. For example, while at wave 2, 50 individuals and at wave 3, 80 individuals are reported to have retired it does not follow that 30 (80 minus 50) individuals retired between waves 2 and 3. The number retired will be greater than 30 owing to some retirees being lost to follow-up between the two waves. Accordingly, by wave 12 more than 241 men will have made the transition to retirement.

Descriptive statistics for our stock sample of men are summarized in Table 7.3 for the sample as a whole and broken down by retirement status. These were obtained using the following commands:

• local summvars “retired hlltyes sah4ex sah4gvg sah4fair sah4vpp shlltyes ssah4ex ssah4gvg ssah4fair

Table 7.2 Labour market status by wave

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Attrition		68	116	146	176	199	209	222	235	248	262	278
Self-employed	162	138	124	105	97	89	76	68	53	46	49	38
Employed	479	349	270	228	193	169	147	129	104	93	78	62
Unemployed		19	20	23	13	10	9	5	6	5	0	2
Retired		50	80	105	125	141	168	187	214	226	233	241
LT sick, disabled		11	23	25	27	28	25	22	19	13	14	10
Other		2	2	2	1	1	3	1	4	5	1	1
Death		4	6	7	9	4	4	7	6	5	4	9
Total	641	641	641	641	641	641	641	641	641	641	641	641
In work*	641	487	394	333	290	258	223	197	157	139	127	100

* Employed and self-employed.

ssah4vpp m2lnhinc HseOwn HseMort HseRent HseAuthAss marcoup deghdeg hndalev ocse everppenr everemppr privcomp0 civlocgov jbsecto0 lspjb"

- summ 'summvars'
- summ 'summvars' if retired==0
- summ 'summvars' if retired==1

Table 7.3 Descriptive statistics

	All	*Pre-Retirement*	*Post-Retirement*
Retired	.324	0	1
Own Health			
hlltyes	.156	.127	.216
sah4ex	.238	.257	.197
sah4gvg	.486	.485	.488
sah4fair	.213	.200	.240
sah4vpp	.064	.058	.075
Spousal Health			
shlltyes	.180	.166	.207
ssah4ex	.156	.171	.126
ssah4gvg	.431	.437	.419
ssah4fair	.191	.192	.191
ssah4vpp	.085	.084	.088
Covariates			
m2lnhinc	9.76	9.76	9.76
HseOwn	.522	.421	.732

HseMort	.320	.415	.122
HseRent	.046	.054	.027
HseAuthAss	.112	.109	.118
marcoup	.867	.886	.827
deghdeg	.084	.087	.078
hndalev	.180	.188	.164
ocse	.217	.214	.223
everppenr	.402	.454	.274
everemppr	.539	.527	.563
privcomp0	.503	.488	.534
civlocgov0	.137	.123	.167
jbsecto0	.100	.098	.105
lspjb	.429	.551	.219

The majority of individuals report SAH status as good or very good. It is notable that the reporting of fair and poor/very poor health increases from pre- to post-retirement while the reporting of excellent health is lower post-retirement. Similarly, the reporting of heath limitations increases post-retirement. Interestingly, the proportion of men whose partner, should they have one, reported fair or poor/very poor SAH is roughly the same pre- and post-retirement. The reporting of health limitations by a partner increases post-compared to pre-retirement.

Of the other variables, the majority of individuals own their house outright (HseOwn); this proportion increases after retirement and is accompanied by a decrease in the proportion of people with an outstanding mortgage. In the rental sector the proportion living in Local Authority rented accommodation (HseAuthAss) increases after retirement. The majority of individuals sampled (52%) do not have an educational qualification. Men are likely to have paid into either a private or occupational pension scheme (94%). 50% of individuals report working within the private sector and approximately 41% of men have a spouse or partner who is in employment.

7.3 THE ENDOGENEITY OF HEALTH

In attempting to identify a causal effect of health on the retirement decision, the use of subjective measures of health has been the focus of much attention (see, for example, Anderson and Burkhauser 1985; Bazzoli 1985; Stern 1989; Bound 1991; Kerkhofs and Lindeboom 1995; Bound *et al.* 1999). However, there are problems in relying on self-reported measures of health status. First, self-reported measures are based on subjective judgements and there is no reason to believe that these judgements are comparable across individuals (see Chapter 4 for further discussion of this issue). Second, self-reported health may not be independent of labour market status. Third, since ill health may represent a legitimate reason for a person of working age to be outside the labour force, respondents not working may cite health problems as a way to rationalize behaviour. Fourth, for individuals for whom the financial rewards of continuing in the labour force

are low there exists a financial incentive to report ill health as a means of obtaining disability benefits, which is often cited as the 'disability route into retirement' (Riphahn 1997; Blundell *et al.* 2002). For example, in a study of social security benefit programmes in the Netherlands, Kerkhofs and Lindeboom (1995) show that recipients of disability insurance systematically overstated their health problems. However, in general, empirical studies on the role of health on retirement provide mixed conclusions about the endogeneity of SAH and the extent of the bias created through measurement error.

To deal with the potential difficulties arising from the use of a selfreported measure of health we follow the approach set out in Bound (1991) and implemented in Bound *et al.* (1999) and subsequently adopted by Disney *et al.* (2006) and Au *et al.* (2005). This involves estimating a model of SAH as a function of more objective measures of health to define a latent 'health stock' variable. This health stock variable is then used as an indicator of health in the model of retirement. The idea of constructing a health stock is analogous to using objective measures of health to instrument the endogenous and potentially error-ridden SAH variable.

Consider the aspect of health that affects an individual's decision to retire, h_{it}^{R}, to be a function of objective and specific measures of health, z_{it}, such that:

$$h_{it}^{R} = z_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1,2,\ldots,n; \quad t = 1,2,\ldots T_i \tag{7.1}$$

where ε_{it} is a time varying error term uncorrelated with z_{it}. We do not directly observe h_{it}^{R} but instead observe a measure of SAH, h_{it}^{S}. We can specify the latent counterpart to h_{it}^{S} as h_{it}^{*} such that:

$$h_{it}^{*} = h_{it}^{R} + \eta_{it} \quad i = 1,2,\ldots,n; \quad t = 1,2,\ldots T_i \tag{7.2}$$

where η_{it} represents measurement error in the mapping of h_{it}^{*} to h_{it}^{R} and is uncorrelated with h_{it}^{R}. Substituting (7.1) into (7.2) gives:

$$h_{it}^{*} = z_{it}\boldsymbol{\beta} + \varepsilon_{it} + \eta_{it} = z_{it}\boldsymbol{\beta} + \nu_{it} \quad i = 1,2,\ldots,n; \quad t = 1,2,\ldots T_i \tag{7.3}$$

The presence of η_{it} in (7.3) represents measurement error, which may be related to labour market status of the individual, and is the source of the bias that would be obtained if we were to use h_{it}^{*} directly when estimating the impact of health on retirement behaviour. To avoid such bias we use the predicted health stock, $\hat{h}_{it}^{*}$, which is purged of measurement error.

Combining (7.3) with the observation mechanism linking the categorical or dichotomous indicator, h_{it}, to the latent measure of health, h^*_{it}, and assuming a distributional form for v_{it} we can estimate the coefficients, β. For example, in the case of the categorical self-assessed measure of health the observation mechanism can be expressed as:

$$h_{it} = k \quad if \quad \mu_{k-1} < h^*_{it} \leq \mu_k, \; k = 1, \ldots, m$$

where $\mu_0 = -\infty$, $\mu_k \leq \mu_{k+l}$, $\mu m = \infty$. Assuming that v_{it} is normally distributed, model (7.3) can be estimated as an ordered probit using maximum likelihood. The predicted values for the health stock can then be used in our retirement model.

Computing the health stock

We construct our health stock variable using a pooled ordered probit model by regressing our measure of self-assessed health (hlstatc4) onto a set of 'objective' health measures. For the latter we use the set of variables on health problems (hlparms to hlpothr). The following command is used:

• oprobit hlstatc4 hlparms hlpsee hlphear hlpskin hlpchest hlpheart hlpstom hlpdiab hlpanx hlpalch hlpepil hlpmigr hlpothr

See Table 7.4 (over page).

Table 7.4 Ordered probits for self-assessed health

Ordered probit estimates				Number of obs=		5449
				LR chi2 (13)=		1731.05
				Prob>chi2=		0.0000
Log likelihood=	−5656.06			Pseudo R2=		0.1327
hlstatc4	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hlparms	−.5818711	.0328725	−17.70	0.000	−.6463001	−.5174421
hlpsee	−.4033917	.0700121	−5.76	0.000	−.5406128	−.2661705
hlphear	−.0902711	.0392977	−2.30	0.022	−.1672931	−.0132491
hlpskin	−.2482158	.055881	−4.44	0.000	−.3577406	.138691
hlpchest	−.6621975	.0483275	−13.70	0.000	−.7569175	−.5674774
hlpheart	−.6228053	.0348666	−17.86	0.000	−.6911426	−.5544679
hlpstom	−.6267968	.0627513	−9.99	0.000	.749787	−.5038065
hlpdiab	−.7773897	.0747391	−10.40	0.000	−.9238756	−.6309039
hlpanx	−.6515209	.0763856	−8.53	0.000	−.8012339	−.5018079
hlpalch	−1.286594	.3272145	−3.93	0.000	−1.927922	−.6452649
hlpepil	−.3914404	.1728269	−2.26	0.024	−.7301749	−.0527059
hlpmigr	−.2186143	.0866258	−2.52	0.012	−.3883978	−.0488308

hlpothr	−.8190626	.069569	−11.77	0.000	−.9554153	−.6827098
_cut1	−2.54126	.0417798	(Ancillary parameters)			
_cut2	−1.339949	.0287601				
_cut3	.2179189	.0242035				

To obtain the latent health stock variable, termed sahlat, we use the predict command with the xb option to specify that we require the linear index. We predict the health stock for individuals for whom we have the relevant set of health variables by specifying e (sample).

• predict sahlat if e (sample), xb

The estimated coefficients display the expected negative sign—health problems are associated with lower reporting of self-assessed health. All effects are highly statistically significant. The dominant effect (in terms of the size of the coefficient) is health problems associated with the use of alcohol or drugs, but problems with arms, legs or hands, chest and breathing, heart or blood pressure, stomach or digestion, diabetes, anxiety or depression, and problems reported as other are also notable.

From the ordered probit model we also construct a latent health stock of an individual's spouse or partner, should they have one—slatsah. This allows us to investigate the effect of spousal health on an individual's retirement decision.

Defining a health shock

Of further relevance is whether the transition to retirement is best identified by a 'shock' to an individual's health or by a levels effect, through a slow deterioration in health status. It is often argued that modelling health 'shocks' is a convenient way of eliminating one source of potential endogeneity bias caused through correlation between individual-specific unobserved factors and health (see, for example, Disney *et al.* 2006).

To identify a health shock we include as variables in our duration model of retirement a measure of health lagged one period together with initial period health. By conditioning on initial health we can interpret the estimated coefficient on lagged health as representing a deviation from some underlying health stock and, accordingly, this approach has the advantage of controlling for person-specific unobserved health-related heterogeneity. Lagged health may be more informative about the decision to retire than contemporaneous health simply because transitions take time. That is, it may take time to adjust fully to a health limitation to enable an individual to assess his/her ability to work or to learn whether an employer can or will accommodate a health limitation.

We compute initial period health and health lagged one period as follows:

• sort pid wavenum
• by pid: gen hlltyes0=hlltyes [1]
• by pid: gen sahlat0=sahlat[1]
• by pid: gen lhlltyes=hlltyes [_n−1]
• by pid: gen lsahlat=sahlat [_n−1]

7.4 EMPIRICAL APPROACH TO DURATION MODELLING

Descriptive analysis

Before proceeding to estimating duration models, we describe the pattern of responses using xtdes. To invoke this command we must first define the individual identifier, pid, using the command iis and the cross-section identifier, wavenum, using the command tis as follows:

- sort pid wavenum
- iis pid
- tis wavenum
- xtdes, patterns (20)

which produces:

```
      pid: 10014608, 10020179, …, 19130392                n=      641
wavenum: 1, 2, …, 12                                      T=       12
         Delta(wavenum)=1; (12–1)+1=12
         (pid*wavenum uniquely identifies each observation)
Distribution of T_i:     min     5%     25%     50%     75%     95%     max
                           1      1       4      12      12      12      12
     Freq.   Percent    Cum.   Pattern
-------------------------------------------------------------------------------
       354     55.23   55.23   111111111111
        72     11.23   66.46   1...........
        50      7.80   74.26   11..........
        32      4.99   79.25   1111.......
        31      4.84   84.09   111........
        21      3.28   87.36   11111111111.
        18      2.81   90.17   11111.......
        16      2.50   92.67   1111111.....
        13      2.03   94.70   1111111111..
        12      1.87   96.57   11111111....
        12      1.87   98.44   111111111...
        10      1.56  100.00   111111......
-------------------------------------------------------------------------------
       641    100.00           xxxxxxxxxxxx
```

This clearly shows that we observe only full sequences of responses. That is, the sequences are not interrupted by missing data at a particular wave. Individuals may or may not have retired during the course of a sequence.

Stata's survival time commands

Before proceeding to estimate discrete-time hazard models, we are first required to prepare the dataset in a manner suitable for implementing the suite of commands Stata employs for the analysis of survival time data. This is achieved using Stata's st (survival time) commands. We assume that the dataset is organized in such a way that, for each individual, there are as many rows as there are time intervals at risk of retirement (or in more general applications events such as death), and accordingly each individual contributes *Ti* rows of data, where *Ti* is the number of waves the individual is observed up to and including the wave of retirement, or the wave in which the individual is censored. This corresponds to a standard unbalanced panel data format.

In addition to having a unique identifier for each individual (pid) we require an identifier for each discrete time at which the person is at risk of retirement. This can be generated from our wave identifier (wavenum). We further require a binary variable to indicate the time interval of retirement. For an individual who is observed to be at risk over a number of discrete-time points and is then observed to retire, this variable will be equal to *0* for all points up to retirement (wave 1, …, T_i-1) and equal to *1* for the final observation period (wave T_i). If an individual is censored at time T_i, *0* would be recorded for all time-periods (waves: 1, …, T_i). The variable, retired, is coded in the manner described. We are now able to prepare the data for analysis by ensuring that they are sorted and invoking the stset command as follows:

- sort pid wavenum
- stset wavenum, id (pid) failure (retired==1) origin(wavenum==1)

The option origin specifies that the first wave of the data represents the origin of the first observation period, that is, the beginning of the period during which an individual becomes at risk of retiring.

These commands return the following:

id:	pid
failure event:	retired==1
obs. time interval:	(wavenum[_n−1], wavenum]
exit on or before:	failure
t for analysis:	(time-origin)
origin:	wavenum==1
5468	total obs.
641	obs. end on or before enter ()
1650	obs. begin on or after (first) failure
3177	obs. remaining, representing
569	subjects
314	failures in single failure-

	per-subject data	
3177	total analysis time at risk, at risk from t=	0
	earliest observed entry t=	0
	last observed exit t=	11

Note that _d is a binary variable indicating retirement status, while _t represents the number of time periods a person is at risk of retirement. Given that we have 12 waves of BHPS data and that the first wave has been specified as the origin, we observe 11 periods over which individuals are at risk. We can tabulate the number of retirement events by time period as: tab _t _d, providing:

_t	0-d	1	Total
1	519	50	569
2	427	44	471
3	371	27	398
4	309	37	346
5	270	27	297
6	236	27	263
7	201	25	226
8	165	29	194
9	141	19	160
10	122	15	137
11	102	14	116
Total	2863	314	3177

Similarly, retirement events can be tabulated by health limitations, using tab_d lhlltyes, to return:

	lhlltyes		
_d	0	1	Total
0	2557	305	2862
1	253	61	314
Total	2810	366	3176

The table shows that while 17% of men reporting health limitations retire over the observation period, only 9% reporting no health limitations retire.

Life tables

Further descriptive analysis of the impact of health on the decision to retire can be achieved using life-table methods. Life tables provide an estimate of the survival, failure

or hazard function associated with a categorical variable (in our example, hlltyes). These are obtained using the ltable command. By default, the corresponding output is provided in table form, but it can also be displayed graphically.

An important consideration in the use of ltable is the underlying process leading to the observations of the events of interest. In our example, although we only observe retirement at the end of each observation period (that is, when a BHPS sample member is interviewed), we assume that the underlying survival time (time to retirement) is continuous, but we are unable to observe the exact timing of retirement. In such circumstances, estimates of the underlying hazard rate are derived from assumptions about the shape of the hazard within each time interval and it is common to assume that events occur at a uniform rate within intervals. This is often termed an *actuarial adjustment* and ltable applies this adjustment by default.

The life-table estimate of survival is obtained as follows. Assume that in the interval t_j to t_{j+1}, we observe d_j transitions to retirement, while c_j individuals are censored. Further assume that there are n_j individuals at risk at the beginning of the interval. By assuming that the censoring process is such that the censored survival times occur at a uniform rate across the interval of interest, the average number of individuals who are at risk during the interval is: $n_j^r = n_j - c_j/2.$ Accordingly in the *j-th* interval, the probability of retirement becomes: $d_j/n_j^r.$ The probability that an individual survives (does not retire) beyond time t_k, that is until some time after the start of the *k-th* interval, is:

$$S(t) = \prod_{j=1}^{k} (n_j^r - d_j) \Big/ n_j^r$$

This is the life table or actuarial estimate. The estimated probability of survival to the *beginning* of the first interval is unity, and the probability of survival *within* any interval is constant.

The following command is used to produce the life table for retirement by health limitations (defined at baseline). The test option produces chi-squared tests of the difference between the groups (health limitations versus no health limitations), failure indicates that a cumulative failure table is required (1−survival).

• ltable _t (_d), by (hlltyes0) test tvid (pid) failure

This produces the output shown in Table 7.5.

Table 7.5 Life table for retirement by health limitations

Interval		Beg. Total	Deaths	Lost	Cum. Failure	Std. Error	[95% Conf.	Int.]
hlltyes0 0								
1	2	527	43	45	0.0852	0.0124	0.0639	0.1132
2	3	439	40	27	0.1712	0.0172	0.1404	0.2079
3	4	372	26	24	0.2311	0.0195	0.1954	0.2721

4	5	322	33	12	0.3114	0.0219	0.2707	0.3566
5	6	277	26	7	0.3768	0.0233	0.3331	0.4243
6	7	244	25	9	0.4419	0.0242	0.3959	0.4907
7	8	210	23	6	0.5039	0.0247	0.4565	0.5533
8	9	181	25	4	0.5732	0.0249	0.5250	0.6223
9	10	152	17	3	0.6214	0.0247	0.5733	0.6697
10	11	132	15	5	0.6653	0.0243	0.6176	0.7124
11	12	112	14	98	0.7396	0.0258	0.6881	0.7887
hlltyes0 1								
1	2	42	7	3	0.1728	0.0594	0.0864	0.3287
2	3	32	4	2	0.2796	0.0718	0.1653	0.4485
3	4	26	1	1	0.3078	0.0744	0.1875	0.4790
4	5	24	4	0	0.4232	0.0813	0.2832	0.5971
5	6	20	1	0	0.4520	0.0822	0.3085	0.6249
6	7	19	2	1	0.5113	0.0833	0.3617	0.6807
7	8	16	2	1	0.5743	0.0836	0.4196	0.7383
8	9	13	4	1	0.7105	0.0799	0.5512	0.8532
9	10	8	2	1	0.7877	0.0750	0.6290	0.9113
10	11	5	0	1	0.7877	0.0750	0.6290	0.9113
11	12	4	0	4	0.7877	0.0750	0.6290	0.9113

Likelihood-ratio test statistic of homogeneity (group=hlltyes0) chi2(1)=1.9125977, P=.16667501

Logrank test of homogeneity (group=hlltyes0):

Log-rank test for equality of survivor functions

hlltyes0	Events observed	Events expected
0	287	296.01
1	27	17.99
Total	314	314.00
	chi2 (1)=	4.92
	Pr>chi2=	0.0266

The likelihood ratio test of homogeneity does not reject the null hypothesis that the failure function is equivalent across men who do and do not report health limitations, while the log-rank test for quality rejects the null at the 5% level.

We can graph the output of It able using the graph option, shown below for survival estimates (proportion not retired):

• ltable_t (_d), by (hlltyes0) test tvid (pid) graph title (“Survival estimates”)

This produces the estimates displayed as Figure 7.1.

The figure indicates that men reporting health limitations (graph on the right) appear to be associated with a greater probability of retiring compared to men not reporting limitations. However, the relationship is not clear, given the size of the estimated confidence intervals placed on the point estimates.

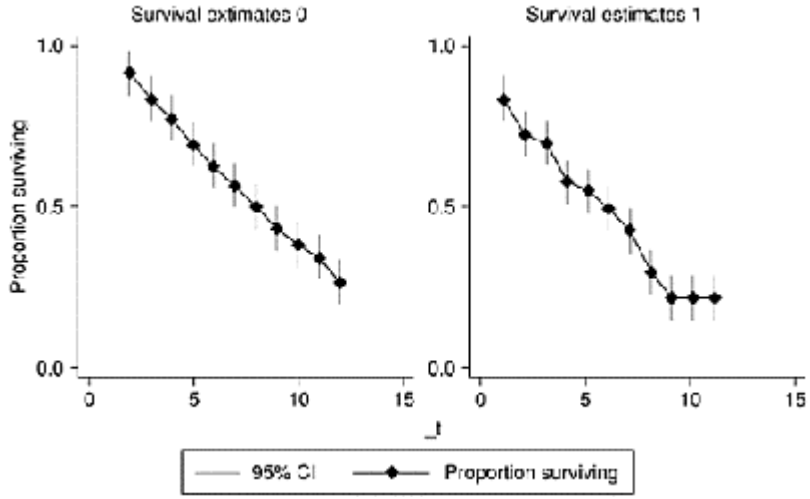

Figure 7.1 Life table estimates of the proportion not retired by health limitations.

7.5
STOCK SAMPLING AND DISCRETE-TIME HAZARD ANALYSIS

The starting point for our analysis is the duration model stock-sampling approach of Jenkins (1995). This method represents the transition to retirement as a discrete-time hazard model, enabling us to estimate the effect of covariates on the probability of retirement. The Jenkins (1995) approach relies on organizing the data so that the unit of analysis is the *time at risk of an event*.

By arranging the data in such a manner, and conditioning on stock sampling—so that time periods prior to selection into the stock sample can be ignored—the estimation of a discrete-time hazard model is simplified to such an extent that estimation methods suitable for a binary outcome (retired versus not retired) may be used. The longitudinal nature of the BHPS dataset together with the use of the stset command has ensured that our data are organized appropriately. Throughout we assume that retirement is an absorbing state such that, at most, there exists a single exit into retirement for each individual.

Adopting the notation of Jenkins (1995), we use data for a stock sample of all individuals who are working at wave 1 *(t=τ)*. At the end of the time period for which we have data, each individual will either still be working (censored duration data, δ_i=0), or will have retired (complete duration data, δ_i=l). $t=\tau+si$ is the year when retirement occurs if δ_j=1 and the final year of our observation period if δ_i=0. Accordingly, each respondent, *i,* contributes s_i years of employment spell data in the interval between the start of the first period and the final wave of observation.

The probability of retiring at each time period, *t,* provides information on the duration distribution, and we define the discrete-time hazard rate as:

$$h_{it}=P[T_i=t|T_i\geq t; x_{it}]$$

where x_{it} is a vector of covariates that may vary with time and T_i is a discrete random variable representing the time at which the end of the spell occurs. The sample likelihood based on stock sampling is conditioned on individuals not having retired at the beginning of the sample time period (wave 1). This is the condition upon which individuals were selected into our sample, implying that all periods prior to the selection period can be ignored. The conditional probability of observing the event history of someone with an uncompleted spell at interview is:

$$prob(T_i > t \mid T_i > \tau - 1) = \prod_{t=\tau}^{\tau+s_i}(1 - h_{it})$$

and the conditional probability of observing the event history of someone completing a spell between the beginning period, τ, and interview is:

$$prob(T_i = t \mid T_i > \tau - 1) = h_{i\tau+s_i}\prod_{t=\tau}^{\tau+s_i-1}(1 - h_{it}) = (h_{i\tau+s_i}/(1 - h_{i\tau+s_i}))\prod_{i=\tau}^{\tau+s_i}(1 - h_{it})$$

Accordingly, the corresponding log-likelihood of observing the event history data for the whole sample is:

$$\log L = \sum_{i=1}^{n}\delta_i \log(h_{i\tau+s_i}/(1 - h_{i\tau+s_i})) + \sum_{i=1}^{n}\sum_{t=\tau}^{\tau+s_i}\log(1 - h_{it})$$

The log-likelihood can be simplified by defining: y_{it}=1 if $t=\tau+s_i$ and δ_i=1, y_{it}=0 otherwise. Accordingly, for stayers y_{it}=0 for *all* spell periods, while for exiters y_{it}=0 for *all* periods *except* the exit period. At exit, y_{it}=1. The likelihood can then be expressed as:

$$\log L = \sum_{i=1}^{n}\sum_{t=\tau}^{\tau+s_i} y_{it} \log(h_{it}/(1 - h_{it})) + \sum_{i=1}^{n}\sum_{t=\tau}^{\tau+s_i}\log(1 - h_{it})$$

To complete the specification of the likelihood, an expression for the hazard rate, h_{it}, is required. We specify a complementary log-log hazard rate, which is the discrete-time counterpart of the hazard for an underlying continuous-time proportional hazards model (Prentice and Gloeckler 1978):

$$h_{it}=1-\exp(-\exp(x_{it}\beta+\theta-\exp(t)))$$

where $\theta(t)$ is an appropriately specified baseline hazard. The model can be generalized to account for unobserved heterogeneity uncorrelated with the explanatory variables (Narendranathan and Stewart 1993).

Estimation in Stata

Stata does not have a suite of built-in commands to estimate directly discrete-time hazard models. However, these models can be easily estimated using existing commands. Before applying any of the commands described, the data must be arranged in the panel data form described for the estimation of life tables. The reader is referred to Jenkins (1995) for details on how to transform discrete-time survival data collected on individuals and stored in wide format (one row per individual) to data stored in long format (multiple rows per individual, one for each discrete time at which observations are made).

There are a number of ways of estimating discrete-time hazard models using either existing Stata commands or, alternatively, by downloading the stb program pgmhaz8 developed by Jenkins (1998). Estimation is via maximum likelihood and follows the form of ML estimation of a binary dependent variable since at any discrete point in time we observe whether an individual has retired (Y=1) or not (Y=0). Accordingly, models for binary dependent variables, for example probit, logistic and complementary log-log models, can be used to model discrete-time hazard functions. Jenkins (1995) provides an intuitive overview of these methods.

For models without frailty (unobserved heterogeneity) we use the complementary log-log command, cloglog. Prior to estimating the model we need to define variables to summarize the pattern of duration dependence. These variables will be a function of time. A commonly used form in continuous-time duration analysis is the Weibull baseline hazard, $\theta(t)= \log(t)$, which can be computed as: gen Int=In (_t). Greater flexibility in duration dependence can be achieved using a piece-wise constant specification where a dummy variable is included in the hazard model to represent each of the discrete time periods under observation. Within each time period duration dependence is assumed constant. This leads to a semiparametric form for the hazard model analogous to Cox's model for continuous-time duration analysis and can be computed using the following routine:

```
• forvalues j =1/11{
        gen t'j'=0
        recede t' j'0=1 if_t=='j'
        }
```

Before proceeding to estimate hazard functions it is useful to delete missing observations from the data stored in memory on the binary retirement variable _d.

• drop if _d==.

Defining the local macro for the set of regressors of interest we can then estimate the hazard function as follows:

• local survars "lhlltyes hlltyes0 age5054 age5559 age6064 m2lnhinc lHseMort lHseAuthAss lHseRent marcoup deghdeg hndalev ocse everppenr everemppr privcomp0 civlocgov0 jbsecto0 lspjb lshlltyes NorthW NorthE SouthW London Midland Scot Wales"
• cloglog _d 'survars' t2−t11, nolog

where _d is the stset variable representing the retirement event. Notice that we have only specified ten time period dummy variables. This is due to the model containing a constant. We could alternatively suppress the constant (noconstant) and specify t1−t11.

The complementary log-log model is within the class of generalized linear models and alternatively can be estimated using Stata's glm command by specifying a complementary log-log link function together with a binomial density: glm _d 'survars' t2−t11, f(bin) 1(cloglog). Using either the cloglog or glm command produces the results in Table 7.6.

Table 7.6 Discrete-time hazard model—no heterogeneity

Complementary log-log regression				Number of obs=		3006
				Zero outcomes=		2734
				Nonzero outcomes=		272
				LR chi2 (38)		360.18
Log likelihood=		−732.71182		Prob>chi2=		0.0000
_d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lhlltyes	.7898271	.1891964	4.17	0.000	.419009	1.160645
hlltyes0	−.3514558	.2632217	−1.34	0.182	−.8673609	1644493
age5054	−2.106317	.4667896	−4.51	0.000	−3.021207	−1.191426
age5559	−1.368255	.2837399	−4.82	0.000	−1.924375	−.8121348
age6064	−.622649	.2535709	−2.46	0.014	−1.119639	−.1256593
age6569	1.043338	.2387245	4.37	0.000	.5754464	1.511229
m2lnhinc	.2808799	.1490468	1.88	0.059	−.0112464	.5730063
lHseMort	−.2060234	.1533145	−1.34	0.179	−.5065142	.0944675
lHseAuthAss	−.1958628	.2145793	−0.91	0.361	−.6164306	2247049
lHseRent	.0435043	.3166615	0.14	0.891	−.5771408	.6641494
marcoup	−.1998969	.2053826	−0.97	0.330	−.6024394	.2026455

deghdeg	−.4142449	.2475186	−1.67	0.094	−.8993725	.0708826
hndalev	−.0797967	.1841427	−0.43	0.665	−.4407097	2811163
ocse	−.0354776	.1801024	−0.20	0.844	−.3884718	3175166
everppenr	−.5558624	.149295	−3.72	0.000	−.8484751	−.2632496
everemppr	.2085984	.1714759	1.22	0.224	−.1274882	5446851
privcomp0	.6864013	.2010634	3.41	0.001	.2923243	1.080478
civlocgov0	1.088949	.238184	4.57	0.000	.6221169	1.555781
jbsecto0	.6856081	.2609122	2.63	0.009	.1742296	1.196987
lspjb	−.3442063	.141413	−2.43	0.015	−.6213706	−.067042
shlltyes	.1289923	.1708559	0.75	0.450	−.205879	4638637
NorthW	−.0150303	.2382563	−0.06	0.950	−.482004	.4519434
NorthE	.3609449	.2054653	1.76	0.079	−.0417597	.7636496
SouthW	−.0361969	.2333262	−0.16	0.877	−.493508	.4211141
London	−.7571608	.307065	−2.47	0.014	−1.358997	−.1553244
Midland	.1238692	.1945952	0.64	0.524	−.2575304	.5052687
Scot	−.250298	.2749037	−0.91	0.363	−.7890993	.2885033
Wales	.2467317	.2767376	0.89	0.373	−.295664	7891275
t2	1.092652	.2888905	3.78	0.000	.5264375	1.658867
t3	.7552746	.3159243	2.39	0.017	.1360743	1.374475
t4	1.208533	.3009155	4.02	0.000	.61875	1.798317
t5	.9261063	.3274428	2.83	0.005	.2843302	1.567882
t6	1.001793	.3221965	3.11	0.002	.3702993	1.633286
t7	1.234166	.3279245	3.76	0.000	.5914456	1.876886
t8	1.230638	.3271135	3.76	0.000	.5895071	1.871768
t9	.8756357	.3515165	2.49	0.013	.1866759	1.564595
t10	.6970452	.3739023	1.86	0.062	−.0357899	1.42988
t11	.6281944	.3872601	1.62	0.105	−.1308214	1.38721
_cons	−5.552784	1.502955	−3.69	0.000	−8.498523	−2.607045

The alternative to estimating discrete-time hazard models is provided by pgmhaz8 (Jenkins 1997). This command is not built in to stata and has to be downloaded as a stb file. This is very much recommended, as a useful feature of this command is that the estimation procedure automatically incorporates frailty (unobserved heterogeneity). The pgmhaz8 routine for models without frailty is essentially the glm command with a complementary log-log link function and a binomial density function for _d. This is estimated using iterative, re weighted least squares (using the irls option) to maximize the deviance rather than the default of maximization of the log-likelihood. pgmhaz8 is implemented as follows:

• pgmhaz8 ‘survars’ t2-t11, i (pid) d (_d) s (_t) nolog

In many health economic applications it is desirable to fit models that take into account unobserved heterogeneity. For ordinary linear regression analysis the consequences of ignoring unobserved heterogeneity are not serious if the heterogeneity is independent of the sets of regressors. In that case, the conditional mean is unchanged and unobserved heterogeneity is absorbed into the error term.

In duration models, which are non-linear, the treatment of unobserved heterogeneity (frailty) causes more concern. Evidence on the effects of ignoring frailty (where it exists) relates mainly to experience with continuous-time duration models and may lead to:

1 Over-estimation of negative duration dependence and under-estimation of positive duration dependence. This has the effect of exaggerating the rate of failure for individuals with a high unobserved heterogeneity effect and underestimating the rate for failure for individuals with a low effect.
2 Under-estimation of the 'true' effects of positive relationships between regressors and duration and an over-estimation of the effect of negative relationships. This is due to the proportionality assumption (that the regressors act proportionally on the underlying hazard function) being attenuated by unobserved heterogeneity.

The extent of the problems caused by unobserved heterogeneity will vary from application to application and will, in part, depend on the specification of the hazard function as well as the chosen unobserved heterogeneity distribution. Where the baseline hazard function is specified flexibly using piece-wise constants, it has been suggested that the impact of disregarding frailty (where the true model suggests its existence) is diminished.

We can incorporate unobserved heterogeneity into our discrete-time hazard model by using either the panel data command: xtcloglog or the pgmhaz8 command, xtcloglog is the panel data equivalent of the cloglog command, and estimates models with unobserved heterogeneity assumed to be normally distributed and constant over time. The results are shown in Table 7.7.

• xtcloglog d 'survars' t2−t11, i(pid) nolog

Table 7.7 Complementary log-log model with frailty

Random-effects complementary log-log model				Number of obs=		3006
Group variable (i): pid				Number of groups=		519
Random effects u_i ~ Gaussian				Obs per group: min=		1
				avg=		5.8
				max=		11
				Wald chi2 (38)=		79.84
Log likelihood=		−714.94324		Prob>chi2=		0.0001
_d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lhlltyes	1.537091	.4072678	3.77	0.000	.738861	2.335321
hlltyes0	−.1248035	.5677018	−0.22	0.826	−1.237479	9878715
age5054	−4.63428	1.059487	−4.37	0.000	−6.710836	−2.557724
age5559	−3.908309	.9137826	−4.28	0.000	−5.69929	−2.117328

age6064	−2.774493	.7549103	−3.68	0.000	−4.25409	−1.294896
age6569	6182654	.4378626	1.41	0.158	−.2399295	1.47646
m2lnhinc	5637165	.3325576	1.70	0.090	−.0880845	1.215518
lHseMort	0749787	.2809341	−0.27	0.790	6255995	4756421
lHseAuthAss	2672034	.4599591	−0.58	0.561	−1.168707	6342999
lHseRent	2764604	.6350249	−0.44	0.663	−1.521086	9681655
marcoup	−.5557142	.4367013	−1.27	0.203	−1.411633	.3002047
deghdeg	−1.121728	.5533454	−2.03	0.043	−2.206265	−.0371913
hndalev	−.1976392	.4094911	−0.48	0.629	−1.000227	6049485
ocse	−.0229894	.3920396	−0.06	0.953	−.7913729	.7453941
everppenr	−1.444642	.4529306	−3.19	0.001	−2.332369	−. 5569143
everemppr	6245727	.416578	1.50	0.134	−.1919052	1.441051
privcomp0	1.274275	.5390658	2.36	0.018	2177256	2.330825
civlocgov0	2.415549	.7045045	3.43	0.001	1.034745	3.796352
jbsecto0	1.213121	.6547056	1.85	0.064	−.0700778	2.496321
lspjb	−.5014095	.2502372	−2.00	0.045	−.9918653	−.0109537
shlltyes	.185817	.2702776	0.69	0.492	−.3439173	7155514
NorthW	−.0870604	.504639	−0.17	0.863	−1.076135	9020139
NorthE	7370782	.4501812	1.64	0.102	−.1452608	1.619417
SouthW	2319199	.4970898	−0.47	0.641	−1.206198	7423582
London	9936346	.6115755	−1.62	0.104	−2.192301	2050314
Midland	2782426	.4272362	0.65	0.515	−.559125	1.11561
Scot	−.2882809	.6473067	−0.45	0.656	−1.556979	.980417
Wales	1.334644	.6983116	1.91	0.056	−.0340216	2.70331
t2	2.005644	.5194759	3.86	0.000	9874903	3.023798
t3	2.053333	.6622402	3.10	0.002	7553663	3.3513
t4	2.959106	.788933	3.75	0.000	1.412826	4.505387
t5	2.88676	.8870166	3.25	0.001	1.14824	4.625281
t6	3.092823	.932335	3.32	0.001	1.26548	4.920166
t7	3.596729	.9930727	3.62	0.000	1.650342	5.543115
t8	3.910824	1.073766	3.64	0.000	1.80628	6.015367
t9	3.797226	1.158281	3.28	0.001	1.527036	6.067416
t10	3.592827	1.173193	3.06	0.002	1.293411	5.892242
t11	3.560867	1.226619	2.90	0.004	1.156739	5.964996
_cons	−9.018572	3.268855	−2.76	0.006	−15.42541	−2.611733
/lnsig2u	1.633058	.5054133			.6424665	2.62365
sigma_u	2.262633	.5717825			1.378827	3.712944
rho	7568264	.0930164			.5361285	.8933999

Likelihood-ratio test of rho=0 : chibar2 (01)=35.54 Prob>=chibar2 =0.000

The final two rows of Table 7.7 report the standard deviation of the heterogeneity variance (sigma_u) and the proportion of total unexplained variation due to heterogeneity (rho). If the hypothesis that unobserved heterogeneity is zero (rho=0) cannot be rejected, then we may conclude that frailty is unimportant. It is clear from the likelihood ratio test

that unobserved heterogeneity is important. A comparison of the coefficient estimates with those of Table 7.6 appears to confirm expectations reflecting the introduction of frailty. For example, positive coefficients in Table 7.7 are larger in magnitude than the corresponding estimates in Table 7.6. Further the magnitudes of the estimated coefficients on the set of time-period dummies tend to be larger compared to Table 7.6, implying greater duration dependence in the model with unobserved heterogeneity.

Frailty is also incorporated into the pgmhaz8 routine, which assumes gamma distributed unobservable heterogeneity (Meyer 1990). The routine estimates a model with frailty automatically after the non-frailty model results are returned and no additional statement or option is required. Accordingly, one can estimate a model with gamma frailty by invoking the command for pgmhaz8 as specified above. Note that an option (nobeta0) exists to switch off the reporting of the non-frailty model estimates if this is desired.

A further approach to incorporating unobserved heterogeneity in discrete-time duration models is via latent class analysis. The assumption here is that individuals are drawn from a population that consists of a finite number of latent classes, and that each individual in the sample can be regarded as a draw from one of these sub-populations. In many applications it may make more sense to consider heterogeneity as consisting of a small number of classes rather than a continuum. For example, for a two-class latent class model of duration, one may wish to think of one of the two classes as consisting of individuals with low duration dependence and the other class as consisting of individuals with high duration dependence. The attraction of this approach is that it leads to a flexible parameterization of heterogeneity; the drawback is that the latent class approach requires more advanced programming skills than the estimators considered thus far. Cameron and Trivedi (2005) provide a thorough treatment of the latent class models in the context of duration analysis, and Chapter 11 discusses their use with count data.

Table 7.8 reports the results from implementing pgmhaz8. The non-frailty model results have been suppressed to conserve space (using the nobeta0 option) but are equivalent to those reported in Table 7.6. The results of the gamma frailty model support our expectations: the effects of the regression coefficients are generally larger than those of the corresponding non-frailty models and duration dependence is greater with frailty than without. The likelihood ratio test statistic once again clearly rejects the null hypothesis of no frailty.

• pgmhaz8 ‘survars’ t2−t11, i (pid) d (_d) s (_t) nobeta0 nolog

Table 7.8 Discrete-time duration model with gamma distributed frailty

PGM hazard model with gamma frailty				Number of obs=	3006
				LR chi2( )=	.
Log likelihood=	−711.14025			Prob>chi2	.

_d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hazard						
lhlltyes	1.490013	.3188968	4.67	0.000	.8649862	2.115039

hlltyes0	.2169158	.5452602	0.40	0.691	−.8517745	1.285606
age5054	−4.268767	.7605834	−5.61	0.000	−5.759483	−2.778051
age5559	−3.63896	.6448771	−5.64	0.000	−4.902896	−2.375024
age6064	−2.725603	.5803678	−4.70	0.000	−3.863103	−1.588103
age6569	.466289	.419506	1.11	0.266	−.3559277	1.288506
m2lnhinc	.5297521	.2892964	1.83	0.067	−.0372584	1.096763
lHseMort	.1098859	.2580863	0.43	0.670	−.3959539	.6157258
lHseAuthAss	−.2393198	.4102137	−0.58	0.560	−1.043324	.5646843
lHseRent	−.2342656	.5272632	−0.44	0.657	−1.267682	.7991513
marcoup	−.4706233	.367816	−1.28	0.201	−1.191529	.2502828
deghdeg	−1.164578	.5045808	−2.31	0.021	−2.153539	−.1756182
hndalev	−.2987275	.3414148	−0.87	0.382	−.9678882	.3704331
ocse	−.1156191	.3369824	−0.34	0.732	−.7760925	.5448542
everppenr	−1.281668	.3321646	−3.86	0.000	−1.932699	−.6306376
everemppr	.7815473	.371098	2.11	0.035	.0542085	1.508886
privcomp0	1.014746	.3873496	2.62	0.009	.2555548	1.773937
civlocgov0	2.065114	.5442161	3.79	0.000	.9984697	3.131758
jbsecto0	.861207	.522051	1.65	0.099	−.1619941	1.884408
lspjb	−.4409412	.2201509	−2.00	0.045	−.872429	−.0094534
shlltyes	.2011737	.2511582	0.80	0.423	−.2910874	.6934348
NorthW	−.011763	.4584404	−0.03	0.980	−.9102896	.8867636
NorthE	.461706	.3897689	1.18	0.236	−.302227	1.225639
SouthW	−.308968	.4209661	−0.73	0.463	−1.134046	.5161104
London	−.5925821	.4977225	−1.19	0.234	−1.5681	.3829362
Midland	.1279848	.3663108	0.35	0.727	−.5899712	.8459408

_d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Scot	−.3849385	.5578826	−0.69	0.490	−1.478368	7084913
Wales	1.210956	.548533	2.21	0.027	.1358507	2.28606
t2	1.587134	.3535936	4.49	0.000	.8941029	2.280164
t3	1.449285	.408406	3.55	0.000	.6488234	2.249746
t4	2.222219	.4376312	5.08	0.000	1.364477	3.07996
t5	2.063826	.4893528	4.22	0.000	1.104712	3.022939
t6	2.259787	.5111582	4.42	0.000	1.257936	3.261639
t7	2.725451	.5481416	4.97	0.000	1.651114	3.799789
t8	3.070844	.6013995	5.11	0.000	1.892122	4.249565
t9	2.98537	.6747382	4.42	0.000	1.662908	4.307833
t10	2.809595	.7155411	3.93	0.000	1.40716	4.21203
t11	2.826252	.7767925	3.64	0.000	1.303767	4.348737
_cons	−6.85814	2.789004	−2.46	0.014	−12.32449	−1.391792

In varg	

_cons	.7482541	.2535725	2.95	0.003	.2512612	1.245247
Gamma var.	2.113307	.5358766	3.94	0.000	1.285646	3.473793

LR test of Gamma var.=0: chibar2 (01)=43.1431 Prob.>=chibar2 =2.5e−11

The hazard for retiring is negative and highly significant for all age categories except ages 65 to 69. As this age group covers the state retirement age for men, a positive coefficient is expected. There is also a gradient across educational attainment such that higher levels of education are associated with a decreasing hazard of retiring. However, the effect is significant for degree or higher degree (deghdeg) only. These effects are compared with the baseline category of no qualifications. The employment sector variables (measured at the first wave) are positive and significant with the exception of other job sector (jbsecto0) and, accordingly, the hazard of retirement is greater for employees compared to the self-employed (the baseline category). The largest effect is observed for individuals employed within civil and local government (civlocgov0), followed by the private sector (privcomp0) and those employed in other sectors (jbsecto0). We also observe a significant effect of pension entitlements. These variables represent whether an individual has made a contribution into a private pension plan (or an employer has made a contribution on behalf of the individual) during the observation period (everppenr) and whether an individual has been a member of an occupational pension scheme during the observation period (everemppr). The former variable is negative and highly significant, while the latter is positive and significant. The effects of housing tenure (lHseMort, lHseAuthAss, and lHseRent), mean logged household income (m2lnhinc) and marital status are not significant at the 5% level.

Our primary focus is the impact of health on the decision to retire. The results clearly show that men with health limitations have a greater hazard of retirement than men not reporting health limitations. The effect is significant at the 1% level.

We are also interested in the impact on retirement of our measure of latent health stock constructed from the pooled ordered probit regressions reported above. Replacing the health limitation variables by our constructed latent health variables (lsahlat and slatsah) results in the estimates provided in Table 7.9:

• local survars2 “lsahlat sahlat0 age5054 age5559 age6064 m2lnhinc lHseMort lHseAuthAss lHseRent marcoup deghdeg hndalev ocse everppenr everemppr privcomp0 civlocgov0 jbsecto0 lspjb slatsah NorthW NorthE SouthW London Midland Scot Wales”

• pgmhaz8 ‘survars2’ t2−t11, i (pid) d(_d) s(_t) nobeta0 nolog

Table 7.9 Discrete-time duration models with latent self-assessed health

PGM hazard model with gamma frailty				Number of obs=	2969
				LR chi2()=	.
Log likelihood=	−712.8288			Prob>chi2=	.
_d	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
hazard					
lsahlat	−.4543396	.1855174	−2.45	0.014	−.8179471 −.0907321

sahlat0	−.0413213	.2506539	−0.16	0.869	−.532594	.4499514
age5054	−3.897328	.7613803	−5.12	0.000	−5.389606	−2.40505
age5559	−3.265921	.652976	−5.00	0.000	−4.545731	−1.986112
age6064	−2.434739	.6057314	−4.02	0.000	−3.621951	−1.247528
age6569	.5813453	.4204195	1.38	0.167	−.2426617	1.405352
m2lnhinc	.4628763	.2785395	1.66	0.097	−.083051	1.008804
lHseMort	.1996237	.2593611	0.77	0.441	−.3087147	7079622
lHseAuthAss	−.1922852	.3959609	−0.49	0.627	−.9683544	5837839
lHseRent	−.267584	.517262	−0.52	0.605	−1.281399	7462308
marcoup	−.4528357	.3783881	−1.20	0.231	−1.194463	.2887914
deghdeg	−1.021895	.4841391	−2.11	0.035	−1.97079	−.0730002
hndalev	−.2984315	.3281383	−0.91	0.363	−.9415707	3447077
ocse	−.1156887	.3290904	−0.35	0.725	−.760694	.5293165
everppenr	−1.240574	.3398719	−3.65	0.000	−1.906711	−.5744373
everemppr	.7065038	.3640348	1.94	0.052	−.0069914	1.419999
privcomp0	.9312187	.3690667	2.52	0.012	.2078612	1.654576
civlocgov0	1.95283	.5309529	3.68	0.000	.9121816	2.993479
jbsecto0	.8787732	.507918	1.73	0.084	−.1167278	1.874274
lspjb	−.5156572	.2145697	−2.40	0.016	−.936206	−.0951084
slatsah	−.0776676	.1503423	−0.52	0.605	−.372333	2169978
NorthW	−.0684027	.4421731	−0.15	0.877	−.935046	7982406
NorthE	.3720781	.3719562	1.00	0.317	−.3569426	1.101099
SouthW	−.3392873	.4110734	−0.83	0.409	−1.144976	4664018

_d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
London	−.7592523	.4981452	−1.52	0.127	−1.735599	.2170943
Midland	2693735	.3494171	0.77	0.441	−.4154714	.9542184
Scot	4425297	.5466936	−0.81	0.418	−1.514029	6289701
Wales	1.139039	.51959	2.19	0.028	.1206618	2.157417
t2	1.504789	.346837	4.34	0.000	.825001	2.184577
t3	1.441112	.4055616	3.55	0.000	.6462255	2.235998
t4	2.130764	.4411201	4.83	0.000	1.266185	2.995344
t5	1.922376	.4840749	3.97	0.000	.9736066	2.871145
t6	2.197616	.5080531	4.33	0.000	1.201851	3.193382
t7	2.629921	.5599493	4.70	0.000	1.532441	3.727402
t8	3.026526	.6169751	4.91	0.000	1.817277	4.235775
t9	2.856774	.6894079	4.14	0.000	1.505559	4.207989
t10	2.695137	.7455741	3.61	0.000	1.233839	4.156436
t11	2.604082	.7918544	3.29	0.001	1.052075	4.156088
_cons	−6.444829	2.699997	−2.39	0.017	−11.73673	−1.152932
ln varg						

_cons	.652314	.3012917	2.17	0.030	.0617931	1.242835
Gamma var.	1.919979	.5784737	3.32	0.001	1.063742	3.465424

LR test of Gamma var.=0 : chibar2 (01)=31.0316 Prob.>=chibar2= 1.3e−08

Qualitatively the results of the non-health variables are the same as those of the corresponding model with health limitations. Again, health is shown to have a significant impact on the timing of retirement, although, unlike health limitations, latent health is significant at only the 5% level. The coefficient is negative owing to the latent health stock increasing in good health.

7.6 OVERVIEW

The primary focus of this chapter is the role of health in determining retirement behaviours. To this end we consider the role of an objective measure of health limitations (hll It yes) and a measure of underlying latent health stock (latsah) constructed from the results of a pooled ordered probit model of self-assessed health on specific health problems. This provides a means of purging self-assessed health of measurement error. Both these variables are lagged one period to avoid problems of simultaneity. We also condition on the first period's health status so that the estimated effect of lagged health can be interpreted as a health shock. Further we consider the health of a respondent's spouse or partner. Clearly, this can only be defined should a respondent have a spouse or partner and therefore needs to be interpreted alongside the estimated effect of the marital status variable (marcoup).

For health limitations we observe a large, positive and highly significant effect. This implies that the hazard of retiring is greater for individuals experiencing a shock to health that leads to a health limitation. For our constructed measure of underlying latent health we observe a negative and significant coefficient. The latent health scale is increasing in health so that the negative coefficient implies that the retirement hazard increases as health decreases. Again this is interpreted as a shock to health. For both models, the estimated coefficients on spousal health are not significant and, accordingly, for men, there is insufficient evidence that the decision to retire is a function of spousal health.

Part IV
Panel data

8
Health and wages

8.1 INTRODUCTION

To illustrate the use of a range of linear panel data estimators we present an empirical model of the impact of health on wage rates using data from the British Household Panel Survey (BHPS). While a great deal of research effort has been placed on determining the existence and extent of a causal effect of income on health, comparatively little research, particularly on developed economies, has investigated the reverse effect of health on income, or as in the example presented here, the effect of health on wage rates. There are a number of reasons why health may impact on wages in a developed economy. First, increases in health are assumed to lead to increases in productivity, which, in turn, should be reflected in an increased wage rate. Second, apart from their direct effects, an employer may perceive health to be correlated with unobservable attributes of an individual which affect productivity and accordingly offer higher wages to healthier employees. Third, irrespective of actual productivity, employers may discriminate against unhealthy individuals.

The example is of interest as it allows us to estimate and compare a number of linear panel data estimators. These range from pooled OLS estimates, through random and fixed effects estimators to instrumental variable Generalized Least Squares estimators that attempt to account for the potential endogenous relationship between health and wage rates. The last estimators are of particular interest as they rely on instruments that are internal to the model. The example is based on Contoyannis and Rice (2001), where further details of the methods and approaches to estimation can be found. Additional relevant reading can be found in Baltagi (2005; Chapter 7).

8.2 BHPS SAMPLE AND VARIABLES

To illustrate the methods we draw on the first six waves of the BHPS. The BHPS has been described in earlier chapters and does not require further elaboration here.

Our population of interest consists of individuals who were in employment in each of the six waves and, importantly, those who gave responses from which we were able to construct an average hourly wage. We attempt to abstract from issues of labour supply, and confine our analysis to the impact of health status on labour productivity, as proxied by average hourly wages. This is likely to underestimate the full effect of health status on *expected* wages, as those individuals who leave the labour force are likely to have poorer health compared to individuals who continue in employment.

We define a sub-sample of data consisting of sample members who were in either full-time or part-time employment in each of the six waves and who provided valid

responses to the variables used in our model. This leads to a balanced sample of 1,625 individuals, consisting of 833 males and 792 females.

Wage rates

The BHPS does not contain an hourly wage variable and so we constructed hourly wages as follows. First, we divided usual gross monthly pay including overtime (using BHPS variable paygu) derived from the main job of an individual by the number of hours worked per month in their main job, again including overtime (derived from BHPS variables jbhrs and jbot). We obtained the hourly wage in a secondary job (j2has indicates a secondary job, j2pay is the pay and j2hrs are the hours worked) analogously and constructed an overall average wage by taking a weighted average of the hourly wage in the main and secondary jobs with weights equal to the proportions of total working time spent in their main and secondary jobs. Using this procedure we obtained a measure of 'maximum average' productivity; those individuals with relatively low wages are more likely to supplement their income with another job, which may be more highly paid, while those who receive relatively high average wages in their main job should be, *ceteris paribus,* less likely to seek a second job. The Stata code to perform these calculations is as follows:

- /* hours per month, main job */
- gen hrsmthn =jbhrs*4.33+jbot*4.33
- /* wage rate, normal job */
- gen wagenorm =paygu/hrsmthn
- /* wages, 2nd job */
- gen wagejb2=0
- replace wagejb2=(j2pay/j2hrs) if j2has==1
- /* proportion of hours, normal job */
- gen propnorm =hrsmthn/ (hrsmthn+j 2hrs)
- gen propothr =j2hrs/ (hrsmthn+j 2hrs) if j2hrs>0;
- gen wage =(propnorm*wagenorm)+(propothr*wagejb2)
- gen lnwage=ln (wage)

Our model of wages and our approach to estimation rely on specifying time-varying and time-invariant regressors. For the instrumental-variables approaches we employ we are further required to partition each of the set of regressors into exogenous and endogenous component sets. We discuss time-varying and time-invariant regressors in turn.

Time-varying regressors

Of particular relevance to this study are the BHPS survey instruments on health status. We use self-assessed health, which is defined by a response to the question: 'Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?' From the responses to this question we create three dummy variables coded to one if an individual has excellent health (sahex), has good

health (sahgd), or has fair or worse than fair health (sahfp). Note that in our sample the category representing poor and very poor health contained less than 4% of all responses and hence was combined with the category representing fair health. It is hypothesized that increasing health has a positive relationship with wages and as such we expect that the coefficient on excellent and good health will be positive with a larger coefficient on excellent health.

Dummy variables for self-assessed health are created as follows:

- recede hlstat −9 −1=.
- tab hlstat, gen(sahdm)
- ren sahdm1 sahex
- ren sahdm2 sahgd
- ren sahdm3 sahf
- ren sahdm4 sahp
- ren sahdm5 sahvp
- gen sahfp=sahf+sahp+sahvp

We further make use of the responses to the General Health Questionnaire (GHQ) which is contained in the BHPS. The GHQ was originally developed as a screening instrument for psychiatric illness but is often used as an indicator of subjective well-being. There are 12 individual elements to the shortened GHQ: concentration, sleep loss due to worry, perception of role, capability in decision making, whether constantly under strain, perception of problems in overcoming difficulties, enjoyment of day-to-day activities, ability to face problems, loss of confidence, self-worth, general happiness, and whether suffering depression or unhappiness. The respondent is asked to indicate on a four-point ordinal scale how they have *recently* felt with respect to the item in question. A Likert scale is then used to obtain an overall score by summing the responses to each question. We use a composite measure derived from the results of this questionnaire that is increasing in ill health (hlghq1). We expect the coefficient on this variable to be negative. We further hypothesize that health status is endogenous in our model of wages.

We construct two variables to represent union status. The first is a binary variable which equals one if the individual has a recognized workplace union that covers pay and conditions for the type of job in which the individual is employed, and the individual is not a member of this union, and zero otherwise (covnon). The second union variable takes a value one if an individual is a member of this union and zero otherwise (covmem). Following Hildreth (1999) we hypothesize that the impact of unionization is positive, with the effect being larger for members than non-members.

A further variable is constructed indicating whether an individual has undertaken any training or education related to their current employment. Given the expectation that training will not immediately impact on wages we include the lagged value of this variable (ljtrain). We hypothesize that this variable will have a positive coefficient.

We allow for a quadratic function of age and experience by including both the levels of these variables and their squares (age, agesqrd, exp, expsqrd). Age should capture general labour market experience and tenure effects. Experience is calculated as the number of years for which an individual has been doing the same job with their current employer. Conditional on age, this variable captures the effect of within-job tenure and

specific (on-the-job) training. We expect positive coefficients for the levels of each of these variables with their effects declining over the life cycle, leading to a concave function in both experience and age (see Mincer 1974).

To account for the possible geographical segmentation of wages, we include a series of regional dummy variables. We also include a binary variable to indicate workforce sector to distinguish between the public and private sectors (jobpriv). It is possible that this variable is endogenous in wages (Disney and Gosling 1998). We further include a measure of the number of employees at the individual's place of work (jbsize) (see Harkness 1996).

We include indicators of marital status (widow, divsep, nvrmar) to capture household economies of scale and productivity effects that are not captured by other variables. Further we include a variable that measures the number of children aged between 0 and 4 years of age (kids04). Previous work has found a positive and significant coefficient for the presence of children in the household for men and a negative and significant coefficient for women (Harkness 1996). We also include a vector of binary variables to indicate occupational status (prof, ma nag, skillnm, skllm). We assumed these variables to be endogenous given the likelihood of selection into job types on the basis of unobserved characteristics which also impact on wages. Finally, we include a vector of time dummies to control for aggregate productivity effects and inflation.

Time-invariant regressors

Ethnic status is included as an exogenous time-invariant variable, coded one if the respondent is white and zero otherwise (white). Previous work has found a gradient in wages across educational attainment and as such we include indicators of the highest academic qualification achieved (Harkness 1996). Responses are categorized into one of the following: degree or higher degree (deg), Higher National Diploma or equivalent (hndct), 'A' levels or equivalent (alevel), or 'O' levels or equivalent (ocse). The baseline category consists of respondents with no formal qualifications. In order to reduce the demands on the data, we use only the indicator of whether an individual has a degree or higher degree when utilizing the instrumental variable estimators. Our expectation is that educational attainment is endogenous in wages, which is consistent with previous research (e.g. Hausman and Taylor 1981; Cornwell and Rupert 1988; Baltagi and Khanti-Akom 1990).

Table 8.1 presents the variables used in the analysis together with their respective definitions.

Table 8.1 Variable labels and definitions

Label	*Definition*
wage	Average hourly wage
age	Age in years
exp	Duration of spell in current job in years
jbsize	Number of employees at workplace
SouthW	Regional indicator: 1=lives in Southwest
London	Regional indicator: 1=lives in London

Midland	Regional indicator: 1=lives in Midlands
NorthW	Regional indicator: 1=lives in Northwest
NorthE	Regional indicator: 1=lives in Northeast
Scot	Regional indicator: 1=lives in Scotland
Wales	Regional indicator: 1=lives in Wales
covmem	Unionization indicator: 1=Covered union member
covnon	Unionization indicator: 1=Covered non-member
jobpriv	Sector indicator: 1=Employed in the private sector
ljtrain	Training indicator: 1=Received education or training related to current employment in the previous period
widow	Marital status indicator: 1=Widowed
divep	Marital status indicator: 1=Divorced or separated
nvrmar	Marital status indicator: 1=Never married
kids04	Number of children in the household aged 0–4
white	Ethnicity indicator: 1=White
deg	Education indicator: 1=Highest academic qualification is degree or higher degree
ocse	Education indicator: 1=Highest academic qualification is O level/CSE
alevel	Education indicator: 1=Highest academic qualification is A level
hndct	Education indicator: 1=Highest academic qualification is HND or equivalent
hlghq1	General Health Questionnaire: Likert Scale score
sahex	Health Indicator: 1=Self-Assessed health reported as excellent
sahgd	Health Indicator: 1=Self-Assessed health reported as good
prof	Occupation Indicator: 1=Professional
manag	Occupation Indicator: 1=Managerial
skllnm	Occupation Indicator: 1=Skilled non-Manual
skllm	Occupation Indicator: 1=Skilled Manual
jobpt	Employment Indicator: 1=Part-time employee

Descriptive statistics

To allow for heterogeneity in coefficients across genders we split the sample by men and women. In the following we only consider the analysis for men. The reader is referred to Contoyannis and Rice (2001) for results for women. Summary statistics for the full sample (which includes both full-time and part-time workers) of men are presented in Table 8.2. Note that less than 2% of men work part-time. Summary statistics were produced as follows:

• drop if male ~=1

• summ wage lnwage age exp jbsize SouthW London Midland NorthW NorthE Scot Wales covmem covnon jobpriv ljtrain widow divsep nvrmar kids04 white deg ocse alevel hndct hlghq1 sahex sahgd prof manag skllnm skllm jobpt

Table 8.2 Summary statistics for full sample of observations

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	4165	8.194638	4.380659	1.018966	55.85513
lnwage	4165	1.993639	.4588185	.0187881	4.022761
age	4165	39.20144	10.08049	17	73
exp	4165	5.990876	6.508984	0	44
jbsize	4165	298.7509	326.8744	1.5	1000
SouthW	4165	.0979592	.2972951	0	1
London	4165	.0965186	.2953366	0	1
Midland	4165	.1687875	.3746091	0	1
NorthW	4165	.1054022	.3071078	0	1
NorthE	4165	.1567827	.3636394	0	1
Scot	4165	.0821128	.2745695	0	1
Wales	4165	.0533013	.2246607	0	1
covmem	4165	.4328932	.4955357	0	1
covnon	4165	.1623049	.3687746	0	1
jobpriv	4165	.7310924	.443445	0	1
ljtrain	4165	.4055222	.4910518	0	1
widow	4165	.0031212	.0557876	0	1
divsep	4165	.0456182	.2086808	0	1
nvrmar	4165	.1601441	.3667836	0	1
kids04	4165	.2110444	.4840124	0	3
white	4165	.9759904	.1530973	0	1
deg	4165	.15006	.3571731	0	1
ocse	4165	.3229292	.467652	0	1
alevel	4165	.2340936	.4234818	0	1
hndct	4165	.0804322	.2719938	0	1
hlghq1	4165	10.15534	4.483037	0	36
sahex	4165	.3082833	.4618397	0	1

Variable	Obs	Mean	Std. Dev.	Min	Max
sahgd	4165	.5246098	.499454	0	1
prof	4165	.0888355	.2845404	0	1
manag	4165	.3361345	.4724422	0	1
skllnm	4165	.1476591	.3548043	0	1
skllm	4165	.2953181	.4562404	0	1
jobpt	4165	.0127251	.112099	0	1

8.3 EMPIRICAL MODEL AND ESTIMATION

We specify a typical Mincerian wage function such that the natural logarithm of wages is a function of individual-level socioeconomic variables that are either time-varying or time-invariant. This can be represented as follows:

$$w_{it}=x_{it}\beta+Zi\gamma+\alpha i+\eta it,\ i=1,2,\ldots,N,\ t=1,2,\ldots,T \tag{8.1}$$

In equation (8.1) i indexes individuals, while t indexes time periods (waves of the BHPS). w_{it} represents the logarithm of hourly wages, x_{it} is a 1×K vector of time-varying regressors including age, work experience and health. z_i is a 1×G vector of time-invariant regressors including qualifications and ethnicity. β and γ are suitably comformed vectors of parameters. a_i is an individual-specific and time-invariant error component, assumed to be normally distributed with zero mean and variance, $\sigma^2_{\eta\cdot}$ Similarly, η_{it} is a classical mean zero disturbance, assumed to be distributed as $\sigma^2_{\eta\cdot}$ We further assume that η_{it} is uncorrelated with the regressors and the individual specific effects, a_i. The effects, a_i may be correlated with all or part of the vectors x and z. For the instrumental variables estimators that we employ we partition the vectors x and z into exogenous and endogenous components (refer to equation (8.2)).

The approach to estimation is as follows. First, assuming that the error disturbances are uncorrelated with the regressors, we estimate the model by OLS. Under this assumption OLS will be unbiased and consistent. However, we note that the parameters estimates will be inefficient as OLS ignores the fact that we have panel data of repeated cross-sections, and hence errors are correlated within individuals. We use the following commands to produce the results presented in Table 8.3.

• local regvars “lnwage age agesqrd exp expsqrd jbsize covmem covnon jobpriv ljtrain widow divsep nvrmar kids04 hlghq1 sahex sahgd prof manag skllnm skllm white deg SouthW London Midland NorthW NorthE Scot Wales yr9293 yr9394 yr9495 yr9596”

• reg ‘regvars’

Table 8.3 OLS on full sample of observations

Source	SS	df	MS	Number of obs=	4165
				F(33, 4131)=	96.64
Model	381.906058	33	11.5729108	Prob>F=	0.0000
Residual	494.675966	4131	.119747268	R-squared=	0.4357
				Adj R-squared=	0.4312
Total	876.582024	4164	.210514415	Root MSE=	.34605

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0373693	.0039577	9.44	0.000	.02961	.0451286
agesqrd	−.0382722	.0047281	−8.09	0.000	.0475418	−.0290027
exp	.0087803	.0021496	4.08	0.000	.004566	.0129946
expsqrd	−.0322945	.0081515	−3.96	0.000	.0482759	−.0163131
jbsize	.0001552	.0000172	9.05	0.000	.0001216	.0001889
covmem	.1117475	.0142312	7.85	0.000	.0838467	.1396483
covnon	.0100673	.0166159	0.61	0.545	.0225089	.0426435
jobpriv	.0147592	.0143509	1.03	0.304	.0133763	.0428947
ljtrain	.044326	.0113303	3.91	0.000	.0221126	.0665394
widow	.0854187	.097766	0.87	0.382	.1062554	.2770928
divsep	−.0876587	.0262608	−3.34	0.001	.1391439	−.0361734
nvrmar	−.073663	.0172395	−4.27	0.000	.1074617	−.0398644
kids04	.0642412	.0121042	5.31	0.000	.0405105	.0879719
hlghq1	−.00124	.0012663	−0.98	0.328	.0037225	.0012426
sahex	.0656882	.0171545	3.83	0.000	.032056	.0993203
sahgd	.0233362	.0154557	1.51	0.131	.0069654	.0536378
prof	.5432044	.0253148	21.46	0.000	.4935738	.592835
manag	.5127876	.019243	26.65	0.000	.475061	.5505143
skllnm	.2903433	.0208847	13.90	0.000	.2493979	.3312886
skllm	.1402549	.0182848	7.67	0.000	.104407	.1761029
white	−.0260942	.0357905	−0.73	0.466	−.0962629	.0440745
deg	.1705886	.0172554	9.89	0.000	.1367587	.2044186
SouthW	−.0718187	.0206931	−3.47	0.001	.1123883	−.0312492
London	.0763798	.0210362	3.63	0.000	.0351375	.1176221
Midland	−.147842	.0174673	−8.46	0.000	.1820874	−.1135966
NorthW	−.0702431	.0201296	−3.49	0.000	.1097079	−.0307782
NorthE	−.0683254	.0178897	−3.82	0.000	.1033988	−.033252
Scot	−.1351816	.0220309	−6.14	0.000	−.178374	−.0919892
Wales	−.0913026	.0262802	−3.47	0.001	.1428259	−.0397792
yr9293	.0240372	.0169829	1.42	0.157	.0092584	.0573327
yr9394	.0678777	.0170392	3.98	0.000	.0344717	.1012838
yr9495	.0925407	.0171145	5.41	0.000	.0589871	.1260943
yr9596	.130982	.0171871	7.62	0.000	.0972859	.164678
_cons	.6758936	.0883556	7.65	0.000	.502669	.8491182

From the OLS results we can see that coefficients on self-assessed general health exhibit the expected positive sign. The coefficient on excellent health is significant at the 1% level, while the estimated coefficient is not significant for good health (both are contrasted against a baseline of fair, poor and very poor health). While the estimated coefficient on psychological health is negative, reflecting an increase in ill health related to a decrease in wages, the coefficient fails to attain statistical significance.

Also of interest are the coefficients on the occupational status variables, with the results clearly showing a gradient associated with increased wages as we move from skilled manual, through skilled non-manual and managerial to professional occupational status. The baseline category represents unskilled, part-skilled and the armed forces. Employees of larger organizations appear to attract higher wages as do employees who are members of a union. Job training exhibits the expected positive coefficient and is significant at the 1% level. As expected, higher qualification (deg) is associated with higher wage rates. Compared with the South-East (baseline category), workers in other regions, with the exception of London, command lower wage rates. Note that the year dummies exhibit a positive gradient, presumably reflecting wage inflation over the period of observation.

The estimated coefficients on age, agesqrd and exp and expsqrd imply the expected significant concave and quadratic relationship with the logarithm of hourly wages. The impact of the number of employees in the workplace also significantly increases wages, as does unionization, with the expected positive differential between those who are union members and those non-members who are covered by union bargaining and negotiation. The coefficient on the private sector dummy is positive but insignificant. The coefficients on the marital status variables suggest that compared to the baseline of married or living with a partner, individuals who are divorced or separated—together with individuals who have never married—tend to have lower wages.

While OLS is consistent under the assumption of no correlation between the regressors and the error terms, we have noted that it is inefficient as our model (8.1) specifies a random effects (RE) structure to the error. We can estimate this model using the xtreg command:

- xtreg ‘regvars’, re i (pid)
- estimate store raneff

Here xt specifies that we wish to estimate a panel data model (cross-sections in time) and the re option specifies a random effects specification of the error disturbance as in (8.1). To indicate that repeated cross-sections are across individuals we include in the command line i(pid), where i represents individuals and pid is the variable name of the personal identification number in the BHPS uniquely assigned to each individual. The results are shown in Table 8.4.

Table 8.4 RE on full sample of observations

Random-effects GLS regression		Number of ob=	4165
Group variable (i): pid		Number of groups=	833
R-sq: within=	0.1244	Obs per group: min=	5
between=	0.4513	avg=	5.0
overall=	0.3923	max=	5
Random effects u_i~Gaussian		Wald chi2(33)=	1142.51
corr (u_i, X)=0 (assumed)		Prob>chi2=	0.0000

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0507045	.0059669	8.50	0.000	.0390095	.0623995
agesqrd	−.0533491	.0071466	−7.46	0.000	−.0673562	−.0393419
exp	.0044801	.0019612	2.28	0.022	.0006362	.008324
expsqrd	−.0205011	.0078083	−2.63	0.009	−.0358051	−.0051971
jbsize	.0000765	.0000178	4.31	0.000	.0000417	.0001114
covmem	.0829266	.0169693	4.89	0.000	.0496674	.1161857
covnon	.0191547	.0161448	1.19	0.235	−.0124885	.0507979
jobpriv	.0171332	.0193016	0.89	0.375	−.0206972	.0549637
ljtrain	.0164079	.0077434	2.12	0.034	.0012311	.0315846
widow	.0034152	.0889423	0.04	0.969	−.1709085	.1777389
divsep	−.0585176	.0316685	−1.85	0.065	−.1205867	.0035515
nvrmar	−.0413766	.0206926	−2.00	0.046	−.0819334	−.0008197
kids04	.0187232	.0105703	1.77	0.077	−.0019941	.0394405
hlghq1	−.0021245	.0010031	−2.12	0.034	−.0040905	−.0001586
sahex	.0277526	.0138943	2.00	0.046	.0005203	.0549849
sahgd	.0128728	.0114762	1.12	0.262	−.0096202	.0353658
prof	.2589516	.0259969	9.96	0.000	.2079986	.3099046
manag	.2419512	.0199481	12.13	0.000	.2028536	.2810488
skllnm	.1542511	.0213966	7.21	0.000	.1123146	.1961876
skllm	.065302	.0164493	3.97	0.000	.033062	.097542
white	−.0090221	.0687363	−0.13	0.896	−.1437427	.1256984
deg	.2955493	.0309714	9.54	0.000	.2348465	.3562521
SouthW	−.0835608	.0372552	−2.24	0.025	−.1565796	−.0105419
London	.0579975	.0356839	1.63	0.104	−.0119417	.1279367
Midland	−.1461243	.0314533	−4.65	0.000	−.2077716	−.0844771
NorthW	−.110799	.0365343	−3.03	0.002	−.182405	−.039193
NorthE	−.0880494	.0329058	−2.68	0.007	−.1525435	−.0235552
Scot	−.166324	.0412882	−4.03	0.000	−.2472475	−.0854006
Wales	−.1048742	.0464365	−2.26	0.024	−.1958882	−.0138602
yr9293	.0274477	.0099184	2.77	0.006	.008008	.0468873

yr9394	.0709793	.010141	7.00	0.000	.0511034	.0908552
yr9495	.099696	.0104665	9.53	0.000	.079182	.1202099
yr9596	.1396777	.0108529	12.87	0.000	.1184065	.1609489
_cons	.622221	.1387916	4.48	0.000	.3501946	.8942475
sigma_u	.27807532					
sigma_e	.19494339					
rho	.67048195			(fraction of variance due to u_i)		

The RE estimates of the self-assessed health variables are smaller than the corresponding OLS estimates. However, the RE estimates of the standard errors are also smaller, resulting in the estimate of excellent health being significant at the 5% level, while good health remains non-significant but exhibits the expected positive coefficient. Interestingly, the estimate of psychological well-being, while still exhibiting a negative coefficient, is significant at the 5% level under the RE estimator. The majority of the other variables retain similar interpretations to the OLS estimates, albeit mostly at a slightly increased level of significance. Note that the majority of unex-plained variation lies at the individual level, rho=0.67, indicating a large degree of unobserved individual heterogeneity in log-wages. Evidence of heterogeneity to this extent provides support for the use of panel-data approaches rather than OLS.

To help motivate the estimators we will use when relaxing the assumption that the regressors are uncorrelated with the individual unobserved effect, it is useful to re-write our model specification (8.1) in the following way:

$$w_{it}=x_{1it}\beta_1+x_{2it}\beta_2+z_{1i}\gamma_1+z_{2i}\gamma_2+\alpha i+\eta_{it},$$
$$i=1,2,\ldots,N,\ t=1,2,\ldots, \qquad (8.2)$$

where x_1 is a $1\times k_1$ vector of exogenous time-varying variables and x_2 is a $1\times k_2$ vector of endogenous variables $(k_1+k_2=K)$. Similarly, z_1 and z_2 are vectors of exogenous and endogenous time-invariant variables of length $1\times g_1$ and $1\times g_2$ $(g_1+g_2=G)$ respectively. The partitioning of x and z into exogenous and endogenous components is based on *a priori* considerations. Throughout η_{it} is assumed to be uncorrelated with the regressors and the individual specific effects, α_i.

An obvious way of estimating a model where we wish to relax the assumption that the regressors are uncorrelated with the individual specific error component, $\alpha_{i,}$ is to use the within-groups (or fixed effects) panel data estimator. A main advantage of this approach is that even in the presence of such correlation, the estimator is consistent. However, it is inefficient, as it dispenses with degrees of freedom—one for each individual in the estimation—and, perhaps more importantly, it does not identify the coeffi-cients γ_1 and γ_2 on the time-invariant variables z_1 and z_2.

To estimate the fixed-effects model we use the same command as we used for the random effects estimator, but specify fe rather than re. We also drop age from the list of variables owing to concerns over colinearity with the set of time dummies.

• local regvars_fe “lnwage agesqrd exp expsqrd jbsize covmem covnon jobpriv ljtrain widow divsep nvrmar kids04 hlghq1 sahex sahgd prof manag skllnm skllm SouthW London Midland NorthW NorthE Scot Wales yr9293 yr9394 yr9495 yr9596”

- xtreg 'regvars_fe', fe i(pid)
- estimate store fixeff

The results are shown in Table 8.5.

Table 8.5 FE on full sample of observations

Fixed-effects (within) regression		Number of obs=	4165
Group variable (i): pid		Number of groups=	833
R-sq: within=	0.1455	Obs per group: min=	5
between=	0.0143	avg=	5.0
overall=	0.0065	max=	5
		F(30,3302)=	18.74
corr (u_i, Xb)=	−0.7347	Prob>F=	0.0000

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agesqrd	−.0481303	.0110108	−4.37	0.000	−.0697189	−.0265416
exp	.0038217	.0020683	1.85	0.065	−.0002335	.0078769
expsqrd	−.0166481	.0084153	−1.98	0.048	−.0331478	−.0001484
jbsize	.0000411	.0000197	2.09	0.037	2.49e−06	.0000797
covmem	.0781241	.0212441	3.68	0.000	.0364712	.1197769
covnon	.023711	.0177723	1.33	0.182	−.0111349	.0585568
jobpriv	.038906	.0273345	1.42	0.155	−.0146884	.0925003
ljtrain	.0063068	.0077064	0.82	0.413	−.0088029	.0214165
widow	−.0339277	.0938112	−0.36	0.718	−.2178617	.1500064
divsep	−.0270047	.0377882	−0.71	0.475	−.1010954	.047086
nvrmar	.0058038	.0252186	0.23	0.818	−.0436419	.0552495
kids04	.0037053	.0110879	0.33	0.738	−.0180345	.0254452
hlghq1	−.0023311	.0010253	−2.27	0.023	−.0043415	−.0003208
sahex	.0137109	.0142684	0.96	0.337	−.0142649	.0416868
sahgd	.0090108	.0115573	0.78	0.436	−.0136495	.0316711
prof	.0849804	.0293286	2.90	0.004	.0274764	.1424844
manag	.0819904	.0228365	3.59	0.000	.0372153	.1267655
skllnm	.0398242	.0241793	1.65	0.100	−.0075837	.0872321
skllm	.0384736	.0172874	2.23	0.026	.0045784	.0723688
SouthW	.1282471	.1028712	1.25	0.213	−.0734507	.3299449
London	−.0244918	.075733	−0.32	0.746	−.1729803	.1239966
Midland	.1212739	.0919738	1.32	0.187	−.0590576	.3016054
NorthW	−.2957556	.104505	−2.83	0.005	−.5006567	−.0908545
NorthE	.1319611	.1263534	1.04	0.296	−.1157777	.3797
Scot	−.1917346	.2015783	−0.95	0.342	−.5869657	.2034965
Wales	−.0445275	.1150586	−0.39	0.699	−.270121	.181066
yr9293	.075379	.0127044	5.93	0.000	.0504696	.1002883

yr9394	.1668088	.0194975	8.56	0.000	.1285804	.2050372
yr9495	.2442047	.0275099	8.88	0.000	.1902666	.2981428
yr9596	.3313773	.0361766	9.16	0.000	.2604465	.4023081
_cons	2.487739	.1770019	14.05	0.000	2.140695	2.834784
sigma_u	.62393645					
sigma_e	.19491991					
rho	.9110821			(fraction of variance due to u_i)		
F test that all u_i=0 :		F(832, 3302)=		12.41	Prob>F=0.0000	

If we turn to the estimates of the health variables we can see that for psychological well-being the FE estimate is very close to the RE estimate. However, the FE estimate is less efficient and, as such, the associated standard error is slightly larger but the parameter estimate still retains statistical significance at the 5% level. Conversely, the parameter estimates on self-assessed excellent and good health are 50% and 30% less than their respective estimates from the RE model. Note that neither parameter is statistically significant at the 5% level using FE estimation.

Also of interest are the occupational status variables. While the gradient remains apparent having accounted for endogeneity using FE, and the parameter estimates lead us to reject the null hypotheses of zero coefficients, the absolute magnitudes are much diminished. The observed difference between the RE and FE estimates suggests that there is positive selection into occupational categories, which may reflect differing time preference, attitudes to risk or other unobserved factors that are positively correlated with wage rates.

Note that ethnicity (white) and educational attainment (deg) are dropped from the FE estimation owing to their being colinear with the unobserved fixed effects.

We may wish to test formally the difference between the parameters obtained from the RE and FE estimators. Under the hypothesis of correct specification of (8.1) and no correlation between x and α_1, the FE estimates of β should be close to the RE results. This can be tested formally using the Hausman test (Hausman 1978). The test statistic is constructed as $M = \boldsymbol{q}'\text{cov}(\boldsymbol{q})^{-1}\boldsymbol{q}$, where $\boldsymbol{q} = \hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}$ and $\text{cov}(\boldsymbol{q}) = \text{cov}\hat{\boldsymbol{\beta}}_{FE} - \text{cov}\hat{\boldsymbol{\beta}}_{RE}$, is asymptotically distributed under H_0 as χ^2_K. Significant differences between the two vectors suggest mis-specification and point to the use of fixed effects or instrumental-variables techniques to overcome endogeneity. However, as noted above, the Hausman test compares only the coefficients on the time-varying regressors. Hence, the employment of the instrumental-variables estimators may remain productive even if the null of exogeneity is not rejected—that is, if one believes that one or more of the time-invariant regressors is/are endogenous with respect to the unobserved individual effect. Note that one may wish to subject to the test only those time-varying regressors deemed, *a priori,* to be correlated with α_i, (x_2), as the test has low power when including all regressors.

The commands estimate store randeff and estimate store fixeff used previously store the estimates from the random effects and fixed effects estimations. By calling on these estimates the Hausman test is invoked by typing:

• hausman fixeff raneff

This produces the following set of results:

Note: the rank of the differenced variance matrix (29) does not equal the number of coefficients being tested (30); be sure this is what you expect, or there maybe problems computing the test. Examine the output of your estimators for anything unexpected and possibly consider scaling your variables so that the coefficients are on a similar scale.

	|---Coefficients---|		
	(b) fixeff	(B) raneff	(b−B) Difference	sqrt (diag (V_b−V_B)) S.E.
agesqrd	−.0481303	−.0533491	.0052188	.0083763
exp	.0038217	.0044801	−.0006584	.0006568
expsqrd	−.0166481	−.0205011	.0038529	.003138
jbsize	.0000411	.0000765	−.0000355	8.46e−06
covmem	.0781241	.0829266	−.0048025	.012781
covnon	.023711	.0191547	.0045562	.0074297
jobpriv	.038906	.0171332	.0217727	.0193552
ljtrain	.0063068	.0164079	−.0101011	.
widow	−.0339277	.0034152	−.0373429	.0298298
divsep	−.0270047	−.0585176	.0315129	.0206169
nvrmar	.0058038	−.0413766	.0471804	.014415
kids04	.0037053	.0187232	−.0150179	.0033483
hlghq1	−.0023311	−.0021245	−.0002066	.0002126
sahex	.0137109	.0277526	−.0140416	.0032461
sahgd	.0090108	.0128728	−.003862	.0013669
prof	.0849804	.2589516	−.1739712	.0135766
manag	.0819904	.2419512	−.1599608	.0111165
skllnm	.0398242	.1542511	−.1144269	.0112616
skllm	.0384736	.065302	−.0268284	.0053176
SouthW	.1282471	−.0835608	.2118079	.0958881
London	−.0244918	.0579975	−.0824893	.0667993
Midland	.1212739	−.1461243	.2673982	.0864284
NorthW	−.2957556	−.110799	−.1849566	.0979108
NorthE	.1319611	−.0880494	.2200105	.1219934
Scot	−.1917346	−.166324	−.0254105	.1973046
Wales	−.0445275	−.1048742	.0603467	.1052717
yr9293	.075379	.0274477	.0479313	.007939
yr9394	.1668088	.0709793	.0958295	.0166527
yr9495	.2442047	.099696	.1445088	.025441
yr9596	.3313773	.1396777	.1916996	.0345103

b=consistent under Ho and Ha; obtained from xtreg
B=inconsistent under IIa, efficient under IIo; obtained from xtreg

```
Test: Ho: difference in coefficients not systematic
chi2(29)=(b−B)'[(V_b−V_B)^(−1)](b−B)
=308.47
Prob>chi2=0.0000
(V_b−V_B is not positive definite)
```

A not uncommon problem encountered when using the Hausman test is that for any finite sample we have no reason to believe that the matrix $\text{cov}(q) = \text{cov}\hat{\beta}_{FE} - \text{cov}\hat{\beta}_{RE}$ is positive definite (PD). If it is not PD then inverting the matrix becomes less than straightforward and a standard application of the Hausman test may not lead to a reliable test statistic. An alternative test, which is asymptotically equivalent to the Hausman test, is an augmented regression. There are several forms of this test and the one used here is based on the following regression:

$$w_{it} = \bar{x}_i\lambda_1 + (x_{it} - \bar{x}_i)\lambda_2 + \alpha_i + \eta_{it} \quad i = 1,2,\ldots,N \quad t = 1,2,\ldots,T \tag{8.3}$$

Under the null hypothesis that $\lambda_1=\lambda_2$, (8.3) collapses to the RE model. Rejection of the null suggests an FE estimator (a Mundlak (1978) type specification). Tests of the joint equivalence of the parameter estimates can be obtained using a Wald test. A program to perform this test is provided below and is invoked as a Stata .do file. It can be run by specifying the set of time-varying regressors to be tested along with the dependent variable and personal identifier as global variables as follows:

• global depvar lnwage
• global varlist agesqrd exp expsqrd jbsize covmem covnon jobpriv ljtrain widow divsep nvrmar kids04 hlghq1 sahex sahgd prof manag skllnm skllm SouthW London Midland NorthW NorthE Scot Wales
• global id pid
• do "hausman_alt.do"

The content of this .do file is replicated in Box 8.1.

Box 8.1 .do file for a test of fixed versus random effects estimation

```
/* .do file to estimate an alternative to the Hausman test of fixed versus random effects */
   /* See Baltagi (2005, p67) and Davidson and MacKinnon (1993, p89) */
   /* Based on program of Vince Wiggins' posting on the STATA list archive 9 Feb 2004 */
   local depvar $depvar
   local varlist $varlist
   local id $id
   tokenize 'varlist'
   local i 1
```

```
while “‘‘i’’”!= “”{
        qui by ‘id’: egen double M‘i’=mean (‘‘i’’)
        qui by ‘id’: gen double D‘i’=‘‘i’’ −M‘i’
        local newlist ‘newlist’ M‘i’ D‘i’
        local i=‘i’+1
  }
  xtreg ‘depvar’ ‘newlist’, re i ($id)
  tempname b
  matrix ‘b’=e(b)
  qui test M1=D1, notest /* clear test */
  local i 2
  while “‘‘i’’” !=“” {
        if ‘b’ [1, colnumb( ‘b’, “M‘i’”)] != 0 & /*
        */ ‘b’ [1, colnumb( ‘b’, “D‘i’”)] != 0 {
            qui test M‘i’ =D‘i’, accum notest
        }
        local i=‘i’+1
  }
  test
  drop ‘newlist’
```

This returns the following chi-squared statistic: chi2(26)=582.84; Prob>chi2=0.0000, rejecting the RE specification.

Finally we estimate the model using instrumental-variables procedures suggested by Hausman and Taylor (1981; HT) and Amemiya and MaCurdy (1986; AM). These methods rely on specifying instruments for the endogenous variables x_2 and z_2. The idea is that by finding instruments for the endogenous variables, the HT and AM estimators are at least as precise as the fixed effects estimator but may avoid the inconsistency of the RE estimator. Furthermore, they allow estimation of the time-invariant regressors.

The HT estimator specifies the following instruments:

$$A_{HT} = ((x_{1it} - \bar{x}_{1i}), (x_{2it} - \bar{x}_{2i}), \bar{x}_{1i}, z_1)$$

Accordingly, the parameters β_1 are identified by the instruments $(x_{1it} - \bar{x}_{1i})$, while β_2 are identified by the instruments $(x_{2it} - \bar{x}_{2i})$. γ_1 and γ_2 are identified by z_1 and $\bar{x}_{1i}$, respectively. Hence z_1 act as their own instrument (z_1 are assumed exogenous), while the time-invariant endogenous regressors, z_2, are instrumented by the within-individual means of the exogenous timevarying regressors. For identification we require that $k_1 \geq g_2$.

The resulting estimator has the benefit of allowing the estimation of time-invariant variables while also allowing for some of the time-varying and time-invariant regressors to be correlated with the individual unobserved error component. Since instruments are derived from variables internal to the model, we do not have to search for external instruments that are relevant and valid, something that is often hard to achieve in practice. However, the relevance of the instrument set formed by the HT method may be weak, particularly for the endogenous time-invariant variables. Note, however, that in principle

one could add external exogenous variables to the instrument set should they be available. We do not pursue this option here.

While the HT estimator is both consistent and more efficient than the FE estimator if the model is overidentified and the partition of the variables into exogenous and endogenous factors is correct, it is inconsistent if some of the assumed exogenous variables are correlated with a_i. We can test for this using a Hausman test comparing the results of the FE estimator with the HT estimator. Under the null that the overidentifying conditions are valid, a test statistic analogous to *M* above is, in general, asymptotically distributed as $\chi^2_{k_1-g_2}$.

To implement the Hausman and Taylor estimator in Stata use the following command, taking care to specify the set of assumed endogenous variables using the option endog ():

• xthtaylor 'regvars', endog(hlghq1 sahex sahgd prof manag skllnm skllm deg)

The results are presented in Table 8.6. It is worth noting that the estimates differ slightly from those presented in Contoyannis and Rice (2001). This is owing to the different sample sizes used and the Stata version of the HT and AM estimators employing an estimator of the variance components that differs from that used by Contoyannis and Rice.

Table 8.6 Hausman and Taylor IV estimator on full sample of observations

Hausman-Taylor estimation		Number of obs=	4165
Group variable (i): pid		Number of groups=	833
		Obs per group: min=	5
		avg=	5
		max=	5
Random effects u_i~i.i.d.		Wald chi2(33)=	696.15
		Prob>chi2=	0.0000

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
TVexogenous						
age	.0489733	.0082305	5.95	0.000	.0328417	.0651049
agesqrd	−.0485502	.0094985	−5.11	0.000	−.0671668	−.0299336
exp	.0042259	.0018978	2.23	0.026	.0005062	.0079456
expsqrd	−.0178106	.0076873	−2.32	0.021	−.0328775	−.0027438
jbsize	.0000436	.0000179	2.43	0.015	8.45e−06	.0000787

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
covmem	.0811579	.0188851	4.30	0.000	.0441437	.118172
covnon	.0240216	.0161374	1.49	0.137	−.0076071	.0556504
jobpriv	.0420076	.0238761	1.76	0.079	−.0047886	.0888038
ljtrain	.0073522	.0071123	1.03	0.301	−.0065876	.0212919

widow	−.026912	.0859373	−0.31	0.754	−.195346	.141522
divsep	−.0354284	.0340369	−1.04	0.298	−.1021394	.0312826
nvrmar	−.0045849	.0225685	−0.20	0.839	−.0488184	.0396486
kids04	.0038084	.0101422	0.38	0.707	−.0160699	.0236867
SouthW	.0369901	.0714148	0.52	0.604	−.1029803	.1769605
London	−.006764	.0582383	−0.12	0.908	−.120909	.1073809
Midland	.003599	.0609756	0.06	0.953	−.115911	.1231091
NorthW	−.2086427	.0707734	−2.95	0.003	−.3473561	−.0699294
NorthE	−.0132267	.070732	−0.19	0.852	−.1518589	.1254055
Scot	−.2005994	.0952262	−2.11	0.035	−.3872392	−.0139596
Wales	−.0376191	.08313	−0.45	0.651	−.2005508	.1253126
yr9293	.0266645	.0093332	2.86	0.004	.0083716	.0449573
yr9394	.0690092	.0106304	6.49	0.000	.048174	.0898445
yr9495	.0978933	.012466	7.85	0.000	.0734604	.1223263
yr9596	.1361645	.0146102	9.32	0.000	.1075291	.1647999
TVendogenous						
hlghq1	−.0023573	.0009477	−2.49	0.013	−.0042147	−.0004999
sahex	.0137715	.013198	1.04	0.297	−.012096	.0396391
sahgd	.0093465	.0106911	0.87	0.382	−.0116078	.0303007
prof	.0871921	.0270631	3.22	0.001	.0341494	.1402348
manag	.0843504	.0210682	4.00	0.000	.0430575	.1256432
skllnm	.0415992	.0223248	1.86	0.062	−.0021565	.085355
skllm	.0384621	.0159865	2.41	0.016	.0071291	.0697952
TIexogenous						
white	.1573687	.1910952	0.82	0.410	−.217171	.5319083
TIendogenous						
deg	1.216191	.2594526	4.69	0.000	.7076734	1.724709
_cons	.3718601	.2701572	1.38	0.169	−.1576382	.9013585
sigma_u	.86751264					
sigma_e	.19403442					
rho	.95235631	(fraction of variance due to u_i)				

note: TV refers to time varying; TI refers to time invariant.

Conditional on the validity of the instrument set of HT, it is possible to obtain a potentially more efficient estimator, as suggested by Amemiya and MaCurdy (1986; AM). Instead of using the means of the time-varying exogenous variables to (over) identify the parameters of the time-invariant endogenous regressors, Amemiya and MaCurdy use the level of each regressor at each time period. Let x_1^* be an $NT \times TK$ matrix (N denotes the number of individuals in the panel) where each column contains values of x_{kit} for a single time period, for example the ktth column of

$x_{1t}^{*} = (x_{k1t}, \ldots, x_{k1t}, \ldots, x_{kNt}, \ldots, x_{kNt})$ for each $k \in K$. The construction of this instrument set is best illustrated using an example. Suppose we have two individuals, each of whom we observe for three time periods. Further assume that we have a set of observations that for individual 1 consists of the values (34, 54, 23) and for individual 2 the values (37, 56, 25). The resulting instrument set would be formed from the final three columns of the matrix resulting from the following transformation:

$$\begin{bmatrix} 1 & 1 & 34 \\ 1 & 2 & 54 \\ 1 & 3 & 23 \\ 2 & 1 & 37 \\ 2 & 2 & 56 \\ 2 & 3 & 25 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 34 & 54 & 23 \\ 1 & 2 & 34 & 54 & 23 \\ 1 & 3 & 34 & 54 & 23 \\ 2 & 1 & 37 & 56 & 25 \\ 2 & 2 & 37 & 56 & 25 \\ 2 & 3 & 37 & 56 & 25 \end{bmatrix}$$

This leads to an instrument set for the AM estimator defined as:

$$A_{AM} = ((x_{1it} - \bar{x}_{1i}), (x_{2it} - \bar{x}_{2i}), x_{1i}^{*}, z_1)$$

While Hausman and Taylor use each x_1 variable as two instruments, Amemiya and MaCurdy use each of these variables as (T+1) instruments. Following the same reasoning as used to ascertain the conditions for existence of the HT estimator, it can be seen that the order condition for the AM estimator is $Tk_1 \geq g_2$. Although the AM estimator, if consistent, is no less efficient than the HT estimator, consistency requires a stronger exogeneity assumption. Hausman and Taylor only require that the means of the x_1 variables be uncorrelated with the unobserved effects, α_i, while the AM estimator requires uncorrelatedness at each point in time. The extra instruments add explanatory power to the reduced form model for z_2 if there is variation over time in the correlation of x_1 and z_2. Again, using a Hausman test we are able to test the extra exogeneity assumptions by comparing the HT and AM estimators (for details of this application, see Contoyannis and Rice (2001)).

The AM estimator is implemented in a similar fashion to the HT estimator by using the amacurdy option, but note the requirement to state the variable name that identifies the cross-sectional time periods; t (wavenum) (Table 8.7).

• xthtaylor 'regvars', endog(hlghq1 sahex sahgd prof manag skllnm skllm deg) amacurdy t(wavenum)

Table 8.7 Men: Amemiya and MaCurdy IV estimator on full sample of observations

Amemiya-MaCurdy estimation	Number of obs=	4165	
Group variable (i): pid	Number of groups=	833	
	Obs per group: min=	5	
	avg=	5	
	max=	5	
Random effects u_i~i.i.d.	Wald chi2(33)=	706.71	
	Prob>chi2=	0.0000	

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
TVexogenous						
age	−.0501906	.0081281	6.17	0.000	.0342598	.0661214
agesqrd	−.0509899	.009349	−5.45	0.000	−.0693136	−.0326663
exp	.0040201	.0018758	2.14	0.032	.0003436	.0076965
expsqrd	−.0174739	.0076038	−2.30	0.022	−.0323771	−.0025707
jbsize	.0000448	.0000177	2.53	0.011	.0000101	.0000796
covmem	.0793588	.0186716	4.25	0.000	.0427631	.1159546
covnon	.0237346	.0159644	1.49	0.137	−.007555	.0550242
jobpriv	.0361759	.0235155	1.54	0.124	−.0099136	.0822655
ljtrain	.0079132	.0070329	1.13	0.261	−.0058709	.0216973
widow	−.0268362	.0850174	−0.32	0.752	−.1934672	.1397948
divsep	−.032636	.0336543	−0.97	0.332	−.0985972	.0333253
nvrmar	−.0033698	.0223218	−0.15	0.880	−.0471197	.0403801
kids04	.0043819	.0100316	0.44	0.662	−.0152796	.0240434
SouthW	.0129475	.0700543	0.18	0.853	−.1243564	.1502515
London	.0017632	.0575254	0.03	0.976	−.1109845	.114511
Midland	−.0119789	.0600301	−0.20	0.842	−.1296358	.1056779
NorthW	−.2069535	.0700134	−2.96	0.003	−.3441771	−.0697298
NorthE	−.0204878	.0699192	−0.29	0.770	−.1575269	.1165512
Scot	−.193395	.0941696	−2.05	0.040	−.377964	−.008826
Wales	−.04507	.0821892	−0.55	0.583	−.206158	.1160179
yr9293	.0273557	.0092296	2.96	0.003	.009266	.0454454
yr9394	.0704982	.0105015	6.71	0.000	.0499155	.0910808
yr9495	.1001087	.0123033	8.14	0.000	.0759946	.1242228
yr9596	.1393102	.014404	9.67	0.000	.1110789	.1675414
TVendogenous						
hlghq1	−.0023621	.0009371	−2.52	0.012	−.0041988	−.0005255
sahex	.0137238	.013053	1.05	0.293	−.0118595	.0393072
sahgd	.0092508	.0105751	0.87	0.382	−.0114761	.0299777

prof	.0916372	.0267275	3.43	0.001	.0392523	.1440221
manag	.0891452	.0207941	4.29	0.000	.0483894	.129901
skllnm	.0450761	.0220462	2.04	0.041	.0018664	.0882858
skllm	.039213	.0158023	2.48	0.013	.008241	.0701849
TIexogenous						
white	.0645989	.1857891	0.35	0.728	−.2995411	.4287389

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
TIendogenous						
deg	.7392345	.1832807	4.03	0.000	.380011	1.098458
_cons	.5320254	.2603319	2.04	0.041	.0217842	1.042266
sigma_u	.86751264					
sigma_e	.19403442					
rho	.95235631	(fraction of variance due to u_i)				

note: TV refers to time varying; TI refers to time invariant.

A further refinement to the instrument set was suggested by Breusch, Mizon and Schmidt (1989; BMS) who, following Amemiya and MaCurdy, make greater use of the time-varying endogenous variables by treating x_2 in a manner similar to Amemiya's and MaCurdy's treatment of x_1. We will not pursue this refinement here as at present STATA does not implement the BMS estimator. Contoyannis and Rice (2001) show the efficiency gains from moving from the AM to the BMS estimator for the example of health and wages described in this chapter.

The results from implementing the HT estimator in Table 8.6 can be compared with those from the AM estimator in Table 8.7. Focusing on the endogenous time-varying variables we can see that the coefficients on the health variables retain the expected signs and mimic closely the corresponding estimates observed for the FE estimator. This is the case for both the HT and AM estimates. Similarly, estimates obtained for occupational class reflect those obtained from the FE estimator. However, for all estimates we observe efficiency gains from using the instrumental-variables approach, with greater gains observed for AM over HT.

To aid comparison across the different estimators, the sets of results are replicated in Table 8.8. It is noteworthy that the FE and instrumentalvariables estimates are consistently lower than the RE estimates, with the within-estimate on sahex 50% lower than that obtained using RE. This may indicate positive correlation between the individual effects and self-assessed health, with those individuals that are more productive, or at least able to obtain relatively high wages, having unobserved characteristics that lead to better self-assessed health.

Table 8.8 Men—comparison across estimators

NT=4165 N=833	OLS	RE	FE	HT	AM
age	.037	.051	–	.049	.050
	(.0040)	(.0060)		(.0082)	(.0081)
agesqrd	−.038	−.053	−.048	−.049	−.051
	(.0047)	(.0071)	(.0110)	(.0095)	(.0093)
exp	.009	.004	.004	.004	.004
	(.0021)	(.0020)	(.0021)	(.0019)	(.0019)
expsqrd	−.032	−.021	−.017	−.018	−.017
	(.0082)	(.0078)	(.0084)	(.0077)	(.0076)
jbsize	.00016	.00008	.00004	.00004	.00004
	(.00002)	(.00002)	(.00002)	(.00002)	(.00002)
covmem	.112	.083	.078	.081	.079
	(.0142)	(.0170)	(.0212)	(.0189)	(.0187)
covnon	.010	.019	.024	.024	.024
	(.0166)	(.0161)	(.0178)	(.0161)	(.0160)
jobpriv	.015	.017	.039	.042	.036
	(.0144)	(.0193)	(.0273)	(.0239)	(.0235)
ljtrain	.044	.016	.006	.007	.008
	(.0113)	(.0077)	(.0077)	(.0071)	(.0070)
widow	.085	.003	−.034	−.027	−.027
	(.0978)	(.0889)	(.0938)	(.0859)	(.0850)
divsep	−.088	−.059	−.027	−.035	−.033
	(.0263)	(.0317)	(.0378)	(.0340)	(.0337)
nvrmar	−.074	−.041	.006	−.005	−.003
	(.0172)	(.0207)	(.0252)	(.0226)	(.0223)
kids04	.064	.019	.004	.004	.004
	(.0121)	(.0106)	(.0111)	(.0101)	(.0100)
Hlghq1	−.001	−.002	−.002	−.002	−.002
	(.0013)	(.0010)	(.0010)	(.0009)	(.0009)
sahex	.066	.028	.014	.014	.014
	(.0172)	(.0139)	(.0143)	(.0132)	(.0131)
sahgd	.023	.013	.009	.009	.009
	(.0155)	(.0115)	(.0116)	(.0107)	(.0106)
prof	.543	.259	.085	.087	.092
	(.0253)	(.0260)	(.0293)	(.0271)	(.0267)
manag	.513	.242	.082	.084	.089
	(.0192)	(.0199)	(.0228)	(.0211)	(.0208)

skllnm	.290	.154	.040	.042	.045
	(.0209)	(.0214)	(.0242)	(.0223)	(.0220)
skllm	.140	.065	.038	.038	.039
	(.0183)	(.0164)	(.0173)	(.0160)	(.0158)
white	−.026	−.009	–	.157	.064
	(.0358)	(.0687)		(.1911)	(.1858)
deg	.171	.296		1.216	.739
	(.0173)	(.0310)	–	(.2595)	(.1833)

1. Age was dropped from the within regression owing to perfect colinearity with the year dummies.
2. Constant, year and regional dummies suppressed.
3. Standard errors are given in parentheses.
4. agesqrd=age^2/100, expsqrd=exp^2/100.

The coefficients on deg indicate a rate of return to having a degree of between 0.3 (RE) and 1.2 (HT). Results of a comparable magnitude have been reported elsewhere (Harkness 1996). The coefficient on deg using the HT instrument set is around four times the magnitude of the corresponding RE estimate. This differential diminishes as stronger exogeneity assumptions are employed using the AM estimator. The results would appear to suggest that individuals who obtain a degree or higher degree appear to be compensating for unobserved characteristics that would otherwise reduce their wages. This result has also been obtained by Hausman and Taylor (1981), Cornwell and Rupert (1988), and Baltagi and Khanti-Akom (1990), and can be rationalized by considering a model where schooling or educational attainment is assumed to be endogenously determined, as considered by Griliches (1977).

Our results bear a striking resemblance to those obtained by Cornwell and Rupert and Baltagi and Khanti-Akom. Using data from the Panel Study of Income Dynamics for a similar number of observations and time periods, and an analogous specification, they also found the majority of the efficiency gains from the AM estimator (and indeed the BMS estimator) to be attached to the coefficient of the time-invariant endogenous variables, and the estimated coefficient to gradually approach the GLS estimates. In particular, our results show that the AM estimate of the standard error of the coefficient on deg is around 70% of that for the HT estimator. This result is to be expected; as the additional AM instruments are timeinvariant, the majority of their additional explanatory power will impact on the time-invariant variables.

It can also be seen that, with the exception of the coefficient on deg, the precision of the instrumental-variables estimators differs negligibly from that of RE. Hence, application of the HT and AM estimators allows us to obtain precise estimates, while avoiding the potential bias and inconsistency of the RE estimator.

8.4 OVERVIEW

This chapter considers the effect of self-assessed general and psychological health on hourly wages using longitudinal data from the six waves of the British Household Panel Survey. We employ single-equation fixed effects and random effects instrumental-

variable estimators suggested by Hausman and Taylor (1981) and Amemiya and MaCurdy (1986). Our results show that reduced psychological health reduces hourly wages for men. We also confirm the findings of previous work by Cornwell and Rupert (1988) and Baltagi and Khanti-Akom (1990), which suggested that the majority of the efficiency gains from the use of the instrumentalvariables estimators fall on the time-invariant endogenous variables, in our case academic attainment, and add further support to the hypothesis of a negative correlation between educational attainment and individual characteristics affects wages.

However, some difficulties of interpretation remain. First, while controlling for endogeneity due to correlation between included explanatory variables and the unobservable individual effects, we have not controlled for potential simultaneity. Hence, although our measures of health are, to some extent, predetermined, our estimates may be contaminated by simultaneity bias as found in previous cross-sectional analyses. Second, our analysis has concentrated on the impact of health variation for the employed. Larger effects may be expected were we to extend our analysis to consider selective participation and allow for endogenous labour supply.

9
Modelling the dynamics of health

9.1 INTRODUCTION

Panel data on individual self-reported health can be used to estimate nonlinear models for binary and ordered dependent variables. These can be based on static or on dynamic specifications. This chapter follows a similar structure to the paper by Contoyannis, Jones and Rice (2004), but rather than analysing an ordered categorical measure of self-assessed health the focus is on a binary measure of limiting health problems. To illustrate the methods we use a panel-data model for a binary measure of health applied to data drawn from the British Household Panel Survey (BHPS). The binary variable is based on the question 'does health limit your daily activities?'

As the analysis estimates models that are designed for panel data we begin by specifying the individual (i) and time indexes (t) and using these to sort the data so that observations are listed by waves within individuals:

- iis pid
- tis wavenum
- sort pid wavenum

The dependent variable is based on the BHPS variable hllt. This needs to be checked, by running descriptive statistics and tabulating the raw data, and then recoded to deal with missing values and cases where there was no answer. Note that in the raw data the variable is coded as 1 for 'yes' and 2 for 'no'. This is recoded to the more usual 0/1 scale so that it is recognized as a standard binary variable by the software:

• sum hllt

Variable	Obs	Mean	Std. Dev.	Min	Max
hllt	64741	1.840194	.4104114	−9	2

• tab hllt

hllt	Freq.	Percent	Cum.
missing	20	0.03	0.03
not answered	2	0.00	0.03
yes	10,120	15.63	15.67
no	54,599	84.33	100.00
Total	64,741	100.00	

• gen hprob=hllt
  • recede hprob −9=.
  • recede hprob −1=.
  • recede hprob 2=0
  • sum hprob

Variable	Obs	Mean	Std. Dev.	Min	Max
hprob	64719	.1563683	.363207	0	1

The final summarize command describes the binary variable that is used in the econometric models.

The next command creates the first of a series of globals to provide a shorthand label for the list of regressors that are used in the econometric models. The list includes measures of gender (male), marital status (widowed nvrmar divsep), educational attainment (deghdeg hndalev ocse), household size and composition (hhsize nch04 nch511 nch1218), a cubic function of age (age age2 age3), race (nonwhite), professional group (prof mantech skillmn ptskill unskill armed) and the logarithm of equivalized real income (lninc):

• global xvars “male widowed nvrmar divsep deghdeg hndalev ocse hhsize nch04 nch511 nch1218 age age2 age3 nonwhite prof mantech skillmn ptskill unskill armed lninc”

Recall the following code that was used in Chapter 2 to create indicators of whether observations are in the balanced and unbalanced estimation samples. These variables will be needed again below:

• quietly probit hprob $xvars
  • gen insampm=0
  • recode insampm 0=1 if e (sample)
  • sort pid wavenum
  • gen constant=1
  • by pid: egen Ti=sum (constant) if insampm == 1
  • drop constant
  • sort pid wavenum
  • by pid: gen nextwavem=insampm [_n+1]
  • gen allwavesm=.
  • recede allwavesm .=0 if Ti ~= 8
  • recede allwavesm .=1 if Ti == 8
  • gen numwavesm=.
  • replace numwavesm=Ti

To generalize the Stata code it is helpful to define a global for the dependent variable as well as the regressors. This makes it easier to adapt the code for other applications of the same models by simply changing the name of the dependent variable from hprob to the new dependent variable (yvar) in this command:

• global yvar “hprob”

Before estimating the panel data regressions it is helpful to use xtsum to derive summary statistics that exploit the panel dimension of the dataset and that separate the between-individual (cross-section) and within-individual (time-series) variation in the dependent and independent variables. This is done for the full sample and for the balanced panel (allwavesm==1):

• xtsum $yvar $xvarw

Variable		Mean	Std. Dev.	Min	Max	Observations
hprob	overall	.1563683	.363207	0	1	N=64719
	between		.3065844	0	1	n=10264
	within		.2242743	−.7186317	1.031368	T-bar=6.30544
hprobt 1	overall	.154296	.3612354	0	1	N=57798
	between		.3064186	0	1	n=10264
	within		.2189146	−.7028469	1.011439	T-bar=5.63114
male	overall	.461485	.4985182	0	1	N=64741
	between		.499175	0	1	n=10264
	within		0	.461485	.461485	T-bar=6.30758
widowed	overall	.0881745	.2835507	0	1	N=66323
	between		.2834152	0	1	n=10264
	within		.0879394	−.7868255	.9631745	T-bar=6.46171
nvrmar	overall	.1633672	.3697031	0	1	N=66323
	between		.3589536	0	1	n=10264
	within		.131913	−.7116328	1.038367	T-bar=6.46171
divsep	overall	.0682116	.2521106	0	1	N=66323
	between		.224614	0	1	n=10264
	within		.1215966	−.8067884	.9432116	T-bar=6.46171
deghdeg	overall	.0964536	.2952141	0	1	N=82112
	between		.2952267	0	1	n=10264
	within		0	.0964536	.0964536	T=8

Variable		Mean	Std. Dev.	Min	Max	Observations
hndalev	overall	.2024552	.4018321	0	1	N=82112
	between		.4018492	0	1	n=10264
	within		0	.2024552	.2024552	T=8
ocse	overall	.2724084	.4452016	0	1	N=82112
	between		.4452206	0	1	n=10264
	within		0	.2724084	.2724084	T=8
hhsize	overall	2.788357	1.329707	1	11	N=64741
	between		1.2373	1	10.66667	n=10264
	within		.5415378	−2.711643	9.413357	T-bar=6.30758
nch04	overall	.1443753	.4196944	0	4	N=64741
	between		.3263182	0	2.5	n=10264
	within		.2747743	−1.998482	2.519375	T-bar=6.30758
nch511	overall	.2597736	.6145583	0	6	N=64741
	between		.5198938	0	5.25	n=10264
	within		.3368913	−2.597369	3.402631	T-bar=6.30758
nch1218	overall	.1833151	.4861762	0	4	N=64741
	between		.3828021	0	2.5	n=10264
	within		.3109858	−1.691685	2.808315	T-bar=6.30758
age	overall	46.95723	17.77155	15	100	N=64741
	between		18.34678	16	97	n=10264
	within		2.180959	32.15723	52.24294	T-bar=6.30758
age2	overall	25.20804	18.17837	2.25	100	N=64741
	between		18.86734	2.56	94.09	n=10264
	within		2.165909	10.62004	32.12929	T-bar=6.30758
age3	overall	15.01471	15.53261	.3375	100	N=64741
	between		16.2452	.4096	91.2673	n=10264
	within		1.978015	4.10103	24.99951	T-bar=6.30758
nonwhite	overall	.0619641	.2410919	0	1	N=82112
	between		.2411022	0	1	n=10264
	within		0	.0619641	.0619641	T=8
lninc	overall	9.497943	.6664307	−.1312631	13.12998	N=64101
	between		.5793668	4.692182	12.13122	n=10261
	within		.364328	2.399066	12.70143	T-bar=6.24705
prof	overall	.0342062	.1817598	0	1	N=64579
	between		.1470494	0	1	n=10264
	within		.1034673	−.8407938	.9092062	T-bar=6.2918
mantech	overall	.1843943	.3878084	0	1	N=64579
	between		.322583	0	1	n=10264

	within		.2093004	−.6906057	1.059394	T-bar=6.2918
skillmn	overall	.1229037	.3283292	0	1	N=64579
	between		.2790731	0	1	n=10264
	within		.1849436	−.7520963	.9979037	T-bar=6.2918
ptskill	overall	.0859258	.2802566	0	1	N=64579
	between		.2232924	0	1	n=10264
	within		.1838036	−.7890742	.9609258	T-bar=6.2918
unskill	overall	.0262624	.1599159	0	1	N=64579
	between		.1231267	0	1	n=10264
	within		.1093333	−.8487376	.9012624	T-bar=6.2918

Variable		Mean	Std. Dev.	Min	Max	Observations
armed	overall	.0007588	.0275354	0	1	N=64579
	between		.0301781	0	1	n=10264
	within		.0180176	−.713527	.8757588	T-bar=6.2918
hprobt_1	overall	.1383445	.3452634	0	1	N=82056
	between		.3452781	0	1	n=10257
	within		0	.1383445	.1383445	T=8
mlninc	overall	9.467574	.5751038	4.692182	12.24306	N=82088
	between		.5751283	4.692182	12.24306	n=10261
	within		0	9.467574	9.467574	T=8

• xtsum $yvar $xvarw if allwavesm==1

Variable		Mean	Std. Dev.	Min	Max	Observations
hprob	overall	.1444193	.351518	0	1	N=48560
	between		.2707188	0	1	n=6070
	within		.2242471	−.7305807	1.019419	T=8
hprobt_1	overall	.139727	.3467076	0	1	N=42490
	between		.2699814	0	1	n=6070
	within		.2175466	−.7174159	.9968699	T=7
male	overall	.4494234	.4974406	0	1	N=48560
	between		.4974764	0	1	n=6070
	within		0	.4494234	.4494234	T=8
widowed	overall	.0806219	.2722564	0	1	N=48560
	between		.2571574	0	1	n=6070
	within		.0894603	−.7943781	.9556219	T=8
nvrmar	overall	.143925	.3510173	0	1	N=48560
	between		.3244268	0	1	n=6070
	within		.1340729	−.731075	1.018925	T=8
divsep	overall	.0681837	.2520635	0	1	N=48560
	between		.2211882	0	1	n=6070
	within		.1209083	−.8068163	.9431837	T=8
deghdeg	overall	.1143328	.3182183	0	1	N=48560
	between		.3182412	0	1	n=6070
	within		0	.1143328	.1143328	T=8
hndalev	overall	.2253707	.4178305	0	1	N=48560
	between		.4178606	0	1	n=6070
	within		0	.2253707	.2253707	T=8
ocse	overall	.2866557	.452204	0	1	N=48560
	between		.4522365	0	1	n=6070
	within		0	.2866557	.2866557	T=8
hhsize	overall	2.808979	1.302575	1	10	N=48560
	between		1.184421	1	8.625	n=6070
	within		.5422633	−2.691021	9.433979	T=8
nch04	overall	.149547	.4222539	0	4	N=48560
	between		.3119503	0	2	n=6070
	within		.2846039	−1.600453	2.524547	T=8

Variable		Mean	Std. Dev.	Min	Max	Observations
nch511	overall	.2696664	.6204588	0	4	N=48560
	between		.511909	0	3.375	n=6070
	within		.350651	−1.980334	2.894666	T=8
nch1218	overall	.1843904	.4860323	0	4	N=48560
	between		.3667978	0	2.5	n=6070
	within		.3189141	−1.69061	2.80939	T=8
age	overall	46.87016	17.02794	15	100	N=48560
	between		16.87426	18.5	96.5	n=6070
	within		2.29154	42.74516	51.87016	T=8
age2	overall	24.86757	17.28749	2.25	100	N=48560
	between		17.13723	3.475	93.175	n=6070
	within		2.283685	18.18257	31.69257	T=8
age3	overall	14.55261	14.59312	.3375	100	N=48560
	between		14.4457	.6623	90.0152	n=6070
	within		2.076254	4.973107	24.53741	T=8
nonwhite	overall	.0324547	.1772062	0	1	N=48560
	between		.177219	0	1	n=6070
	within		0	.0324547	.0324547	T=8
lninc	overall	9.528054	.6414354	3.324561	12.9514	N=48560
	between		.5398719	6.994533	12.13122	n=6070
	within		.3464387	4.887086	12.58706	T=8
prof	overall	.0351936	.1842707	0	1	N=48560
	between		.1498151	0	1	n=6070
	within		.1073049	−.8398064	.9101936	T=8
mantech	overall	.1932867	.3948799	0	1	N=48560
	between		.3306547	0	1	n=6070
	within		.2159014	−.6817133	1.068287	T=8
skillmn	overall	.1224465	.3278041	0	1	N=48560
	between		.2686998	0	1	n=6070
	within		.1877934	−.7525535	.9974465	T=8
ptskill	overall	.0852348	.2792336	0	1	N=48560
	between		.2076425	0	1	n=6070
	within		.1867143	−.7897652	.9602348	T=8
unskill	overall	.0272446	.1627972	0	1	N=48560
	between		.1173912	0	1	n=6070
	within		.1128016	−.8477554	.9022446	T=8
armed	overall	.0004119	.0202904	0	1	N=48560
	between		.0120002	0	.625	n=6070

	within		.016362	−.6245881	.8754119	T=8
hprobt1	overall	.1120264	.3154021	0	1	N=48560
	between		.3154249	0	1	n=6070
	within		0	.1120264	.1120264	T=8
mlninc	overall	9.532457	.5330323	7.043408	12.24306	N=48560
	between		.5330707	7.043408	12.24306	n=6070
	within		0	9.532457	9.532457	T=8

Note that for time-invariant variables, such as educational qualifications (deghdeg, hndalev, ocse), there is no within-individual variation. Note also that for most of the time-varying variables there is more betweenthan within-individual variation.

Comparing the results for the full sample with those from the balanced sample (allwavesm==1) helps to reveal any systematic differences in the observable characteristics of the samples. The balanced sample is a little healthier than the full sample, with 0.154 reporting health problems in the full sample and 0.144 in the balanced sample. For many of the observed characteristics the differences in means are small but, for example, the balanced sample is better educated with a higher proportion of university graduates (deghdeg). The issues of non-response and attrition bias are pursued in Chapter 10.

9.2 STATIC MODELS

Our models apply to a binary dependent variable: 'does health limit your daily activities?' There are repeated measurements for each wave (t=1,…,T) for a sample of n individuals (i=1,…,n), and the binary dependent variable y_{it} can be modelled in terms of a continuous latent variable y_{it}^*:

$$y_{it} = 1(y_{it}^* > 0) = 1(\boldsymbol{x}_{it}\boldsymbol{\beta} + u_{it} > 0)$$

where 1(.) is a binary indicator function. The error term u_{it} could be allowed to be freely correlated over time or the correlation structure could be restricted. A common specification is the error components model, which splits the error into a time-invariant individual random effect (RE), α_i, and a time-varying idiosyncratic random error, ε_{it}:

$$y_{it} = 1(y_{it}^* > 0) = 1(\boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} > 0)$$

The idiosyncratic error term could be autocorrelated, for example following an AR(1) process, $\varepsilon_{it}=\rho\varepsilon_{it-1}+\eta_{it}$, or it could be independent over t (giving the random effects model).

Pooled specification

The simplest specification is to proceed as if the u_{it} are independent over t and to use a pooled probit model. This applies the standard cross-section probit estimator, even

though there are repeated observations for the same individual. So, the marginal probability of reporting a health problem at wave t is given by:

$P(y_{it}=1|x_{it})=\Phi[(x_{it}\beta)]$

The log-likelihood for the pooled model implicitly assumes that observations are independent across waves and uses the simple product of these marginal distributions. If, as is likely, observations are in fact correlated within individuals this joint distribution will be mis-specified and hence the estimates will not be maximum likelihood estimates (MLE). However, the marginal distributions for each wave are correctly specified even though the joint distribution across waves is incorrectly specified. The properties of the quasi-maximum likelihood estimator (QMLE), which applies in this case, mean that the pooled probit estimates are consistent even though the log-likelihood function is incorrect. However the conventional ML estimates of the standard errors will not be consistent and these need to be replaced by sandwich estimates that are robust to clustering within-individuals (robust cluster (pid)).

The first results are for the unbalanced panel (Table 9.1):

• dprobit $yvar $xvars, robust cluster (pid)

Table 9.1 Pooled probit model, unbalanced panel

Probit regression, reporting marginal effects — Number of obs= 63918
Wald chi2(22)= 1981.96
Prob>chi2= 0.0000
Log pseudolikelihood=−24044.138 — Pseudo R2= 0.1325
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
male*	.0088301	.0057147	1.55	0.122	.461294	−.00237	.020031
widowed*	−.0033601	.0089857	−0.37	0.711	.089646	−.020972	.014252
nvrmar*	.0203347	.0092692	2.27	0.023	.160659	.002167	.038502
divsep*	.0351431	.0109542	3.45	0.001	.069151	.013673	.056613
deghdeg*	−.0563287	.0078981	−6.02	0.000	.108295	−.071809	−.040849
hndalev*	−.044868	.0067819	−6.08	0.000	.21526	−.05816	−.031576
ocse*	−.0551473	.0062399	−8.15	0.000	.279765	−.067377	−.042917
hhsize	−.0002887	.0029376	−0.10	0.922	2.78962	−.006046	.005469
nch04	−.0265443	.0065124	−4.07	0.000	.145014	−.039308	−.01378
nch511	−.0093574	.0049476	−1.89	0.059	.260115	−.019054	.00034
nch1218	−.0062888	.0053513	−1.18	0.240	.182969	−.016777	.0042
age	.0225194	.0032899	6.86	0.000	46.9753	.016071	.028968
age2	−.0382381	.0064794	−5.91	0.000	25.2308	−.050937	−.025539
age3	.0225046	.0039697	5.67	0.000	15.0369	.014724	.030285
nonwhite*	.0927335	.014889	7.24	0.000	.045527	.063552	.121915

prof*	−.0781306	.0082116	−6.61	0.000	.034263	−.094225	−.062036
mantech*	−.0916388	.0049077	−15.19	0.000	.184142	−.101258	−.08202
skillmn*	−.1020752	.0044425	−16.49	0.000	.12308	−.110782	−.093368
ptskill*	−.0856117	.004574	−13.86	0.000	.086048	−.094577	−.076647
unskill*	−.0852802	.0064642	−8.54	0.000	.026299	−.09795	−.07261
armed*	−.0981824	.0286635	−1.65	0.100	.000767	−.154362	−.042003
lninc	−.0327962	.0034939	−9.38	0.000	9.4974	−.039644	−.025948
obs. P	.1563879						
pred. P	.1233529	(at x-bar)					

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient being 0

The second set of results are for the balanced panel (allwavesm==1) (Table 9.2):

• dprobit $yvar $xvars if allwavesm==1, robust cluster(pid)

Table 9.2 Pooled probit model, balanced panel

Probit regression, reporting marginal effects					Number of obs=		48540
					Wald chi2(21)=		1155.09
					Prob>chi2=		0.0000
Log pseudolikelihood== −17749.875					Pseudo R2=		0.1146
		(standard errors adjusted for clustering on pid)					
hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95%	C.I.]
male*	.0050355	.0066901	0.75	0.451	.449279	−.008077	.018148
widowed*	−.008172	.0106382	−0.75	0.452	.080655	−.029022	.012678
nvrmar*	.0286118	.0112959	2.67	0.008	.143923	.006472	.050751
divsep*	.021309	.0122194	1.83	0.067	.068212	−.002641	.045259
deghdeg*	−.0552503	.0089877	−5.16	0.000	.11438	−.072866	−.037635
hndalev*	−.0390154	.0078207	−4.64	0.000	.225278	−.054344	−.023687
ocse*	−.0485099	.0073163	−6.17	0.000	.286712	−.06285	−.03417
hhsize	−.0019848	.0035145	−0.56	0.572	2.809	−.008873	.004903
nch04	−.0307996	.0074678	−4.12	0.000	.149485	−.045436	−.016163
nch511	−.0158376	.005856	−2.70	0.007	.269633	−.027315	−.00436
nch1218	−.0085733	.0060741	−1.41	0.158	.184446	−.020478	.003332
age	.0241683	.0040737	5.94	0.000	46.8719	.016184	.032153
age2	−.0416048	.008134	−5.12	0.000	24.8699	−.057547	−.025662
age3	.0237872	.0050727	4.69	0.000	14.5549	.013845	.033729
nonwhite*	.0961948	.022156	5.14	0.000	.032365	.05277	.13962
prof*	−.0781057	.0087654	−5.93	0.000	.035208	−.095286	−.060926
mantech*	−.0877245	.0055672	−12.96	0.000	.193366	−.098636	−.076813

skillmn*	−.0957497	.0050526	−13.67	0.000	.122497	−.105653	−.085847
ptskill*	−.0814963	.0050922	−11.83	0.000	.08527	−.091477	−.071516
unskill*	−.0759617	.0075648	−6.78	0.000	.027256	−.090788	−.061135
lninc	−.0388673	.0043018	−9.04	0.000	9.52793	−.047299	−.030436
obs. P	.1444788						
pred. P	.1159175	(at x-bar)					

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient being

Note that the models are estimated using the dprobit command rather than probit. This automatically presents the results as partial effects, evaluated at the sample means of the regressors. Stata computes marginal effects—based on derivatives—for continuous variables, and average effects—based on differences—for discrete regressors. This means that the reported results can be given a quantitative as well as a quantitative interpretation. For example those with university degrees (deghdeg) have estimated partial effects of −0.056 and −0.055, respectively, for the unbalanced and balanced samples—meaning that the probability of reporting a limiting illness is around 0.055 lower for those with degrees than those without qualifications (the reference category), holding other factors in the model constant at their sample means. One note of caution here is that this benchmark may not be very meaningful when the dummy variables relate to mutually exclusive sets of categories such as educational qualifications, marital status and occupational group, where each individual can only belong to one category at a time. This creates a problem of interpretation with the standard approach to computing partial effects.

Correlated effects

In the pooled probit model the individual effect (α_i) is subsumed into the overall error term. The model assumes that the individual effect is independent of the observed regressors, an assumption that will often be questionable in applied work. An approach to dealing with individual effects that are correlated with the regressors is to specify $E(\alpha/x)$ directly. For example, in dealing with a random effects probit model Chamberlain (1980) suggests using:

$$a_i = a'x_i + u_i,\ u_i \sim iid\ N(0, \sigma^2)$$

where $x_i=(x_{i1},\ldots, x_{iT})$, the values of the regressors for every wave of the panel, and $a=(\alpha_1,\ldots,\alpha_T)$. This approach will work with the pooled probit model as well. Then, by substitution, the distribution of y_{it} conditional on x_i but marginal to α_i has the probit form:

$$P(y_{it}=1|x_i)=\Phi[(\beta'x_{it}+\alpha'x_i)]$$

In other words the pooled probit model is augmented by adding x_i. A special case of this approach, associated with earlier work by Mundlak (1978) uses the within-individual means of the regressors rather than separate values for each wave. This can be

implemented in Stata by using egen to create new variables for the within-means. Here we only take the within-mean of log(income) (lninc):

• bypid: egen mlninc=mean (lninc)

Then this is added to the list of regressors:

• global xvarm "male widowed nvrmar divsep deghdeg hndalev ocse hhsize nch04 nch511 nch1218 age age2 age3 nonwhite lninc prof mantech skillmn ptskill unskill armed mlninc"

The new list of regressors is used to re-run the pooled probit models (Tables 9.3 and 9.4):

• dprobit $yvar $xvarm, robust cluster (pid)

Table 9.3 Mundlak specification of pooled probit model, unbalanced panel

Probit regression, reporting marginal effects	Number of obs=	63918
	Wald chi2(23)=	1964.72
	Prob>chi2=	0.0000
Log pseudolikelihood =−23961.178	Pseudo R2=	0.1355

(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95%	C.I.]
male*	.0088981	.005701	1.56	0.118	.461294	−.002276	.020072
widowed*	−.0051568	.0089061	−0.57	0.567	.089646	−.022612	.012299
nvrmar*	.0191472	.0092154	2.15	0.032	.160659	.001085	.037209
divsep*	.0296996	.010759	2.94	0.003	.069151	.008612	.050787
deghdeg*	−.0470572	.0085314	−4.82	0.000	.108295	−.063778	−.030336
hndalev*	−.0377078	.0070659	−4.98	0.000	.21526	−.051557	−.023859
ocse*	−.0501993	.0063795	−7.32	0.000	.279765	−.062703	−.037696
hhsize	−.0002862	.0029289	−0.10	0.922	2.78962	−.006027	.005454
nch04	−.0297788	.0065605	−4.53	0.000	.145014	−.042637	−.016921
nch511	−.0136174	.0050027	−2.72	0.007	.260115	−.023423	−.003812
nch1218	−.0094341	.005367	−1.76	0.079	.182969	−.019953	.001085
age	.0243101	.0033084	7.36	0.000	46.9753	.017826	.030794
age2	−.0414329	.0065081	−6.37	0.000	25.2308	−.054189	−.028677
age3	.0241093	.0039808	6.06	0.000	15.0369	.016307	.031912
nonwhite*	.0886518	.0148611	6.91	0.000	.045527	.059525	.117779
lninc	.0020076	.002974	0.68	0.500	9.4974	−.003821	.007837
prof*	−.0741176	.0085928	−6.16	0.000	.034263	−.090959	−.057276

mantech*	−.0871639	.0050307	−14.25	0.000	.184142	−.097024	−.077304
skillmn*	−.1006049	.0044597	−16.25	0.000	.12308	−.109346	−.091864
ptskill*	−.0849738	.0045661	−13.77	0.000	.086048	−.093923	−.076024
unskill*	−.0856156	.0063264	−8.69	0.000	.026299	−.098015	−.073216
armed*	−.093506	.0317633	−1.53	0.125	.000767	−.155761	−.031251
mlninc	−.0591852	.0063148	−9.32	0.000	9.50075	−.071562	−.046808
obs. P	.1563879						
pred. P	.1225589	(at x-bar)					

(*) dF/dx is for discrete change of dummy variable from 0 to 1 z and P>|z| correspond to the test of the underlying coefficient being 0

• dprobit $yvar $xvarm if allwavesm==1, robust cluster(pid)

Table 9.4 Mundlak specification of pooled probit model, balanced panel

Probit regression, reporting marginal effects — Number of obs= 48540
Wald chi2(22)= 1144.69
Prob>chi2= 0.0000
Log pseudolikelihood=−17678.168 — Pseudo R2= 0.1182
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
male*	.0051167	.0066716	0.77	0.442	.449279	−.007959	.018193
widowed*	−.0097037	.0105401	−0.90	0.370	.080655	−.030362	.010955
nvrmar*	.0282309	.0112409	2.65	0.008	.143923	.006199	.050263
divsep*	.0166732	.0119855	1.45	0.147	.068212	−.006818	.040164
deghdeg*	−.0453601	.0098219	−4.03	0.000	.11438	−.064611	−.026109
hndalev*	−.0310701	.008179	−3.59	0.000	.225278	−.047101	−.01504
ocse*	−.0431943	.0074828	−5.42	0.000	.286712	−.05786	−.028528
hhsize	−.0021173	.0035045	−0.60	0.546	2.809	−.008986	.004751
nch04	−.0341893	.007512	−4.55	0.000	.149485	−.048912	−.019466
nch511	−.0204341	.0059213	−3.45	0.001	.269633	−.03204	−.008829
nch1218	−.0113524	.0060825	−1.87	0.062	.184446	−.023274	.000569
age	.025906	.0040987	6.32	0.000	46.8719	.017873	.033939
age2	−.0446999	.008178	−5.46	0.000	24.8699	−.060728	−.028671
age3	.0253243	.0050951	4.97	0.000	14.5549	.015338	.035311
nonwhite*	.0908373	.0221155	4.85	0.000	.032365	.047492	.134183
lninc	−.0015475	.0035184	−0.44	0.660	9.52793	−.008443	.005348
prof*	−.0748357	.009096	−5.61	0.000	.035208	−.092664	−.057008

mantech*	−.0830761	.005717	−12.08	0.000	.193366	−.094281	−.071871
skillmn*	−.0941188	.0050712	−13.44	0.000	.122497	−.104058	−.08418
ptskill*	−.0807503	.0050789	−11.74	0.000	.08527	−.090705	−.070796
unskill*	−.0765929	.0073535	−6.95	0.000	.027256	−.091006	−.06218
mlninc	−.06339	.0076843	−8.21	0.000	9.53228	−.078451	−.048329
obs. P	.1444788						
pred. P	.1149647	(at x-bar)					

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient being 0

The impact is to dramatically reduce the size and statistical significance of current income (lninc), while the effect of mean income is larger and statistically significant.

Random effects specification

Using the pooled model gives estimates of the coefficients and partial effects that are consistent and robust to clustering within individuals. However, using the error components assumption of the random effects model can provide more efficient estimates and provide information on how much of the random variability in health is attributable to the individual effect. Assuming that α and ε are normally distributed and independent of *x* gives the random effects probit model (REP).In this case a can be integrated out to give the sample log-likelihood function by taking the expectation over all the possible values of a weighted by their probability density:

$$\ln L = \sum_{i=1}^{n}\left\{\ln \int_{-\infty}^{+\infty} \prod_{t=1}^{T} (\boldsymbol{\Phi}[d_{it}(\boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha)])\, f(\alpha) d\alpha\right\}$$

where $d_{it}=2y_{it}-1$.This expression contains an integral which can be approximated by Gauss-Hermite quadrature. Assuming $\alpha \sim N(0,\sigma_\alpha^2)$, the contribution of each individual to the sample likelihood function is:

$$L_i = \int_{-\infty}^{+\infty} (1/\sqrt{2}\pi\sigma_\alpha^2)\, \exp(-\alpha^2/2\sigma_\alpha^2)\ \{g(\alpha)\} d\alpha,$$

$$g(\alpha) = \prod_{t=1}^{T} \boldsymbol{\Phi}[d_{it}(\boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha)]\,.$$

where Use the change of variables, $\alpha = (\sqrt{2\sigma_\alpha^2})z$, to give:

$$L_i = (1/\sqrt{\pi}) \int_{-\infty}^{+\infty} \exp(-z^2) \{g((\sqrt{2\sigma_\alpha^2})z)\} dz$$

$$\int_{-\infty}^{+\infty} \exp(-z^2) f(z) dz,$$

As it takes the generic form this expression is suitable for Gauss-Hermite quadrature and can be approximated as a weighted sum:

$$L_i \approx (1/\sqrt{\pi}) \sum_{j=1}^{m} w_j \, g((\sqrt{2\sigma_\alpha^2})a_j)$$

where the weights (w_j) and abscissae (a_j) are tabulated in standard mathematical references, and m is the number of nodes or quadrature points (see e.g., Butler and Moffitt 1982).

In Stata the random effects probit model is estimated using the xtprobit command. The default is to use 12 quadrature points (m=12). Whether or not this is sufficient can be checked by following up the estimation with the command quadchk (Table 9.5):

• xtprobit $yvar $xvars

Table 9.5 Random effects probit model, unbalanced panel

Random-effects probit regression					Number of obs=	63918
Group variable (i): pid					Number of groups=	10261
Random effects u_i ~ Gaussian					Obs per group: min=	1
					avg=	6.2
					max=	8
					Wald chi2(22)=	2150.42
Log likelihood=−17692.313					Prob>chi2=	0.0000
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	−.0601555	.0446561	−1.35	0.178	−.1476799	.0273689
widowed	−.1192292	.0627647	−1.90	0.057	−.2422458	.0037874
nvrmar	.1085197	.058096	1.87	0.062	−.0053463	.2223857
divsep	.204005	.0599426	3.40	0.001	.0865197	.3214902
deghdeg	−.5885337	.088864	−6.62	0.000	−.762704	−.4143634
hndalev	−.3983224	.0647885	−6.15	0.000	−.5253055	−.2713392
ocse	−.502298	.05855	−8.58	0.000	−.6170539	−.3875422
hhsize	−.0025008	.0179746	−0.14	0.889	−.0377304	.0327288
nch04	−.1226955	.0393269	−3.12	0.002	−.1997747	−.0456162

nch511	−.0483194	.0307243	−1.57	0.116	−.108538	.0118992
nch1218	.0100221	.0319076	0.31	0.753	−.0525155	.0725598
age	.100743	.0223713	4.50	0.000	.056896	.1445899
age2	−.1772154	.0443134	−4.00	0.000	−.2640681	−.0903627
age3	.1343095	.0274418	4.89	0.000	.0805247	.1880944
nonwhite	.7511381	.0906089	8.29	0.000	.5735479	.9287283
prof	−.6264083	.0990585	−6.32	0.000	−.8205593	−.4322573
mantech	−.5932056	.0472102	−12.57	0.000	−.685736	−.5006753
skillmn	−.7016125	.0513351	−13.67	0.000	−.8022275	−.6009976
ptskill	−.5207469	.0509832	−10.21	0.000	−.6206722	−.4208216
unskill	−.4173328	.0799182	−5.22	0.000	−.5739696	−.2606959
armed	−1.219248	.7322593	−1.67	0.096	−2.65445	.2159534
lninc	−.1187486	.0218268	−5.44	0.000	−.1615284	−.0759688
_cons	−2.696962	.3976325	−6.78	0.000	−3.476307	−1.917616
/lnsig2u	.9362017	.029001			.8793607	.9930427
sigma_u	1.596958	.0231567			1.552211	1.642996
rho	.7183318	.0058678			.7066897	.7296885

Likelihood-ratio test of rho=0:chibar2(01)=1.3e+04 Prob>=chibar2= 0.000

The table reports estimates of the coefficients, rather than partial effects, so the magnitudes of these cannot be compared directly with the previous results from the dp rob it command. Note that the estimated value of rho, the intra-class correlation coefficient, is 0.72.This implies that 72% of the unexplained variation in limiting health problems is attributed to the individual effect, suggesting a high degree of persistence. These results use the Stata default of 12 points in the quadrature. The reliability of the approximation provided by this default should be checked: quadchk

• quadchk

	Quadrature check			
	Fitted quadrature 12 points	Comparison quadrature 8 points	Comparison quadrature 16 points	
Log likelihood	−17692.313	−17706.342 −14.029297	−17667.354 24.958984	 Difference
		.00079296	−.00141072	Relative difference
hprob: male	−.06015552	−.05834414 .00181138	−.07129692 −.0111414	 Difference
		−.03011156	.18520986	Relative difference
hprob: widowed	−.11922917	−.11675822 .00247095	−.12695061 −.00772144	 Difference
		−.02072436	.06476131	Relative

				difference
hprob:	.10851969	.10857572	.10764231	
nvrmar		.00005602	−.00087738	Difference
		.00051625	−.00808501	Relative difference
hprob:	.20400496	.20431243	.20233113	
divsep		.00030747	−.00167383	Difference
		.00150718	−.00820487	Relative difference
hprob:	−.58853372	−.58181337	−.60878753	
deghdeg		.00672035	−.02025381	Difference
		−.01141881	.03441401	Relative difference
hprob:	−.39832239	−.3935052	−.40849993	
hndalev		.00481719	−.01017754	Difference
		−.0120937	.02555101	Relative difference
hprob:	−.50229804	−.49792537	−.51837094	
ocse		.00437266	−.0160729	Difference
		−.00870532	.03199874	Relative difference
hprob:	−.00250076	−.00259244	−.0018623	
hhsize		−.00009168	.00063846	Difference
		.03666011	−.2553067	Relative difference
hprob:	−.12269548	−.12256372	−.1234222	
nch04		.00013175	−.00072672	Difference
		−.00107384	.00592293	Relative difference
hprob:	−.04831939	−.04811571	−.04929816	
nch511		.00020368	−.00097877	Difference
		−.0042153	.02025628	Relative difference

		Quadrature check		
	Fitted quadrature 12 points	Comparison quadrature 8 points	Comparison quadrature 16 points	
hprob:	.01002215	.00982595	.01085734	
nch1218		−.0001962	.00083519	Difference
		−.01957643	.08333406	Relative difference
hprob:	.10074299	.10012861	.10074517	
age		−.00061438	2.188e−06	Difference
		−.00609848	.00002172	Relative difference
hprob:	−.17721541	−.17615642	−.17725003	
age2		.00105899	−.00003462	Difference
		−.00597571	.00019537	Relative difference
hprob:	.13430954	.1330986	.13655693	
age3		−.00121094	.00224738	Difference
		−.00901605	.01673287	Relative difference
hprob:	.7511381	.74286356	.78296375	
nonwhite		−.00827454	.03182565	Difference
		−.01101601	.0423699	Relative difference
hprob:	−.62640834	−.62393862	−.63008366	
prof		.00246972	−.00367532	Difference
		−.00394266	.00586729	Relative difference
hprob:	−.59320562	−.59146898	−.59358746	
mantech		.00173663	−.00038184	Difference
		−.00292754	.0006437	Relative difference
hprob:	−.70161252	−.70048718	−.70076168	
skillmn		.00112534	.00085084	Difference
		−.00160393	−.00121269	Relative difference
hprob:	−.5207469	−.52010695	−.51828556	
ptskill		.00063995	.00246134	Difference
		−.00122892	−.00472656	Relative difference
hprob:	−.41733277	−.41951428	−.40753075	
unskill		−.00218151	.00980202	Difference
		.00522727	−.02348729	Relative difference
hprob:	−1.2192484	−1.211436	−1.2608374	
armed		.00781236	−.04158908	Difference
		−.00640752	.03411043	Relative difference

hprob:	−.11874863	−.1193847	−.11569534	
lninc		−.00063608	.00305329	Difference
		.0053565	−.02571217	Relative difference

	Quadrature check			
	Fitted quadrature 12 points	Comparison quadrature 8 points	Comparison quadrature 16 points	
hprob:	−2.6969618	−2.6528693	−2.8231901	
_cons		.04409247	−.1262283	Difference
		−.01634894	.04680389	Relative difference
lnsig2u:	.93620169	.89161455	1.0572874	
_cons		−.04458714	.12108572	Difference
		−.04762557	.12933722	Relative difference

There are noticeable discrepancies in the estimates as the number of quadrature points is increased from 8, through 12, to 16: this is especially so for the estimate of the variance component (lnsig2u). So, from now on we will sacrifice computational speed for the sake of improved accuracy and use a higher value, 24 points, in subsequent estimation of the random effects model (using intp (24)). For example, the model is now applied to the balanced sample (Table 9.6):

• xtprobit $yvar $xvars if allwavesm==1, intp (24)

Table 9.6 Random effects probit model, balanced panel

Random-effects probit regression		Number of obs=	48560
Group variable (i): pid		Number of groups=	6070
Random effects u_i~Gaussian		Obs per group: min=	8
		avg=	8.0
		max=	8
		Wald chi2 (22)=	1052.41
Log likelihood=−12659.028		Prob>chi2=	0.0000

hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	−.1373869	.0601367	−2.28	0.022	−.2552527	−.0195212
widowed	−.2748047	.0810101	−3.39	0.001	−.4335816	−.1160278
nvrmar	.1748432	.0721793	2.42	0.015	.0333744	.3163121
divsep	.1326683	.0739552	1.79	0.073	−.0122813	.2776178
deghdeg	−.6740032	.1140367	−5.91	0.000	−.897511	−.4504955
hndalev	−.402745	.0845301	−4.76	0.000	−.568421	−.237069
ocse	−.5112824	.0776967	−6.58	0.000	−.6635651	−.3589997
hhsize	−.0134927	.0221311	−0.61	0.542	−.0568689	.0298835

nch04	−.1299834	.0467827	−2.78	0.005	−.2216758	−.038291
nch511	−.0929403	.0372722	−2.49	0.013	−.1659925	−.019888
nch1218	−.005101	.0381632	−0.13	0.894	−.0798995	.0696974
age	.1346385	.029159	4.62	0.000	.077488	.1917891
age2	−.2614758	.0582107	−4.49	0.000	−.3755666	−.1473849
age3	.1919592	.0364711	5.26	0.000	.1204771	.2634414
nonwhite	.8712417	.1526795	5.71	0.000	.5719953	1.170488
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
prof	−.6206747	.1201381	−5.17	0.000	−.8561412	−.3852083
mantech	−.5406964	.0546823	−9.89	0.000	−.6478717	−.4335211
skillmn	−.6681279	.0605139	−11.04	0.000	−.7867331	−.5495228
ptskill	−.4780917	.059783	−8.00	0.000	−.5952642	−.3609192
unskill	−.3164638	.0902917	−3.50	0.000	−.4934322	−.1394954
armed	−6.548145	5199.183	−0.00	0.999	−10196.76	10183.66
lninc	−.1536884	.0273579	−5.62	0.000	−.2073088	−.1000679
_cons	−2.740474	.512739	−5.34	0.000	−3.745424	−1.735524
/lnsig2u	1.086426	.0434551			1.001256	1.171597
sigma_u	1.72153	.0374046			1.649757	1.796425
rho	.7477082	.0081974			.7313054	.7634335

Likelihood-ratio test of rho=0: chibar2 (01)=1.0e+04 Prob>=chibar2 =0.000

The random effects probit model has two important limitations: it relies on the assumptions that the error components have a normal distribution and that errors are not correlated with the regressors. One way in which normality can be relaxed is to use a finite mixture model (see Deb 2001). This approach is not pursued here but is presented in the context of models for health care utilization in Chapter 11.

The possibility of correlated effects can be dealt with by using conditional (fixed effects) approaches or by parameterizing the effect. To implement the latter approach, these random effects models are now augmented by the Mundlak specification to allow for individual effects that are correlated with the within-individual means of the regressors (in our case lninc) (Tables 9.7 and 9.8):

• xtprobit $yvar $xvarm, intp(24)

Table 9.7 Mundlak specification of random effects probit model, unbalanced panel

Random-effects probit regression				Number of obs=		63918
Group variable (i) : pid				Number of groups=		10261
Random effects u_i ~ Gaussian				Obs per group: min=		1
				avg=		6.2
				max=		8
				Wald chi2 (23)		=1938.63
Log likelihood= −17618.25				Prob > chi2=		0.0000
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	−.0572653	.047638	−1.20	0.229	−.1506341	.0361036
widowed	−.1533264	.064921	−2.36	0.018	−.2805691	−.0260837
nvrmar	.1047099	.0602443	1.74	0.082	−.0133667	.2227864
divsep	.162142	.0619059	2.62	0.009	.0408087	.2834754
deghdeg	−.383708	.0973599	−3.94	0.000	−.5745299	−.1928861
hndalev	−.2588571	.0705589	−3.67	0.000	−.39715	−.1205642
ocse	−.4204988	.063216	−6.65	0.000	−.5443997	−.2965978
hhsize	−.0056164	.0184637	−0.30	0.761	−.0418045	.0305718
nch04	−.1480549	.0402702	−3.68	0.000	−.226983	−.0691268
nch511	−.0845105	.0317611	−2.66	0.008	−.1467612	−.0222598
nch1218	−.0122348	.032702	−0.37	0.708	−.0763295	.0518599
age	.1233626	.023396	5.27	0.000	.0775073	.1692179
age2	−.2144947	.0462538	−4.64	0.000	−.3051505	−.1238389
age3	.1533843	.028592	5.36	0.000	.0973451	.2094235
nonwhite	.7441638	.096222	7.73	0.000	.5555721	.9327555
lninc	−.0069076	.02497	−0.28	0.782	−.0558479	.0420327
prof	−.5763284	.1020212	−5.65	0.000	−.7762863	−.3763705
mantech	−.5522213	.0485459	−11.38	0.000	−.6473695	−.4570731
skillmn	−.6923852	.052473	−13.20	0.000	−.7952304	−.58954
ptskill	−.5196423	.0519204	−10.01	0.000	−.6214045	−.4178802
unskill	−.4183784	.0813274	−5.14	0.000	−.5777772	−.2589795
armed	−1.129763	.7685715	−1.47	0.142	−2.636135	.3766095
mlninc	−.5211792	.053282	−9.78	0.000	−.62561	−.4167484
_cons	.6361512	.5415711	1.17	0.240	−.4253086	1.697611
/lnsig2u	1.056674	.0361678			.9857864	1.127562
sigma_u	1.696109	.0306723			1.637046	1.757304
rho	.7420544	.0069229			.7282548	.7553886

Likelihood-ratio test of rho=0: chibar2 (01)=1.3e+04 Prob>=chibar2= 0.000

• xtprobit $yvar $xvarm if allwavesm==1, intp (24)

Table 9.8 Mundlak specification of random effects probit model, balanced panel

Random-effects probit regression				Number of obs=		48560
Group variable (i): pid				Number of groups=		6070
Random effects u_i ~ Gaussian				Obs per group: min=		8
				avg =		8.0
				max=		8
				Wald chi2 (23)=		1088.01
Log likelihood=−12619.818				Prob>chi2		= 0.0000
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	−.1192647	.0599106	−1.99	0.047	−.2366872	−.0018421
widowed	−.3011138	.0808099	−3.73	0.000	−.4594983	−.1427294
nvrmar	.1850041	.0719871	2.57	0.010	.043912	.3260961
divsep	.0914589	.0739631	1.24	0.216	−.0535061	.2364239
deghdeg	−.3919638	.1174985	−3.34	0.001	−.6222566	−.1616711
hndalev	−.2111355	.0867028	−2.44	0.015	−.3810699	−.0412012
ocse	−.3913465	.0783682	−4.99	0.000	−.5449454	−.2377476
hhsize	−.0195375	.0221159	−0.88	0.377	−.0628839	.0238089
nch04	−.1551483	.0468387	−3.31	0.001	−.2469505	−.0633462
nch511	−.1323934	.0374751	−3.53	0.000	−.2058433	−.0589435
nch1218	−.0285701	.0382921	−0.75	0.456	−.1036214	.0464811
age	.159061	.0292788	5.43	0.000	.1016757	.2164463
age2	−.3008401	.0583309	−5.16	0.000	−.4151666	−.1865135
age3	.208999	.0364707	5.73	0.000	.1375178	.2804803
nonwhite	.7945646	.1517943	5.23	0.000	.4970532	1.092076
lninc	−.0458109	.0299387	−1.53	0.126	−.1044898	.0128679
prof	−.5693845	.1205951	−4.72	0.000	−.8057465	−.3330226
mantech	−.4969237	.0548426	−9.06	0.000	−.6044133	−.3894341
skillmn	−.6597481	.0603996	−10.92	0.000	−.7781291	−.5413671
ptskill	−.4792066	.0596291	−8.04	0.000	−.5960774	−.3623357
unskill	−.3314323	.0901974	−3.67	0.000	−.508216	−.1546486
armed	−6.478364	5441.711	−0.00	0.999	−10672.04	10659.08
mlninc	−.6156707	.0696439	−8.84	0.000	−.7521702	−.4791711
_cons	1.591711	.7064546	2.25	0.024	.2070855	2.976337
/lnsig2u	1.069177	.0425637			.9857541	1.152601
sigma_u	1.706746	.0363227			1.637019	1.779443
rho	.7444405	.0080977			.7282485	.7599856

Likelihood-ratio test of rho=0: chibar2 (01)=1.0e+04 Prob >= chibar2= 0.000

The outcome matches the results for the pooled models; current income (lninc) is no longer statistically or quantitatively significant but mean income is.

Simulation-based inference

The random-effects probit model only involves a univariate integral. More complex models, for example where the error term ε_{it} is assumed to follow an AR(1) process, lead to sample log-likelihood functions that involve higher order integrals. Monte Carlo simulation techniques can be used to deal with the computational intractability of non-linear models, such as the panel probit model and the multinomial probit. Popular methods of simulation-based inference include classical Maximum Simulated Likelihood (MSL) estimation, and Bayesian Markov Chain Monte Carlo (MCMC) estimation (see Contoyannis, Jones and Leon-Gonzalez (2004) for further details).

Recall that the general version of our model is:

$$y_{it} = 1(y_{it}^* > 0) = 1(\boldsymbol{x_{it}\beta} + u_{it} > 0)$$

This implies that the probability of observing the sequence $y_{i1}...y_{iT}$ for a particular individual is:

$$Prob(y_{i1}, \ldots, y_{iT}) = \int_{ai1}^{bi1} \cdots \int_{aiT}^{biT} f(u_{i1}, \ldots, u_{iT}) du_{iT}, \ldots, du_{i1}$$

with $a_{it}=-x_{it}\beta$, $b_{it}=\infty$ if $y_{it}=1$ and $a_{it}=-\infty$, $b_{it}=-x_{it}\beta$ if $y_{it}=0$. The sample likelihood L is the product of these integrals, L_i, over all n individuals. In certain cases, such as the random-effects probit model, L_i can be evaluated by quadrature. In general, the T-dimensional integral L_i cannot be written in terms of univariate integrals that are easy to evaluate. Gaussian quadrature works well with low dimensions but computational problems arise with higher dimensions. Multivariate quadrature uses the Cartesian product of univariate evaluation points and the number of evaluation points increases exponentially. Instead Monte Carlo (MC) simulation can be used to approximate integrals that are numerically intractable. This includes numerous models derived from the multivariate normal distribution (the panel probit, multinomial and multivariate probit, panel ordered probit and interval regression, panel Tobit, etc.). MC approaches use pseudo-random selection of evaluation points and computational cost rises less rapidly than with quadrature (see Contoyannis, Jones and Leon-Gonzalez (2004) for details).

The conditional logit model

The conditional logit estimator uses the fact that the within-individual sum $\Sigma_t y_{it}$ is a sufficient statistic for a_i (see e.g., Chamberlain 1980). This means that conditioning on $\Sigma_t y_{it}$ allows a consistent estimator for β to be derived, using the logistic function:

$P(y_{it}=1|x_{it}, \alpha_i)=F(x_{it}\beta+\alpha_i)=\exp(x_{it}\beta+\alpha_i/(1+\exp(x_{it}\beta+\alpha_i))$

Then, for example in the case where T=2, it is possible to show that:

$$P[(0,1)|(0,1) \text{ or } (1,0)]=\exp((x_{i2}-x_{il})\beta)/(1+\exp(x_{i2}-x_{il})\beta))$$

This implies that a standard logit model can be applied to differenced data and the individual effect is swept out. In practice, conditioning on those observations that make a transition—(0,1) or (1,0)—and discarding those that do not—(0,0) or (1,1)—means that identification of the models relies on those observations where the dependent variable changes over time.

The conditional logit estimator is implemented by the following commands, first for the unbalanced panel and then for the balanced panel (Tables 9.9 and 9.10). The group command specifies the individual identifier:

• clogit $yvar $xvars, group(pid)

Table 9.9 Conditional logit model, unbalanced panel

Conditional (fixed-effects) logistic regression					Number of obs=	18715
					LRchi2(17)=	738.64
					Prob>chi2=	0.0000
Log likelihood= −6596.2496					Pseudo R2=	0.0530
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
widowed	−.3154207	.1599588	−1.97	0.049	−.6289342	−.0019072
nvrmar	.0193504	.1505804	0.13	0.898	−.2757818	.3144825
divsep	.0326483	.1376459	0.24	0.813	−.2371328	.3024293
hhsize	.0380896	.0402778	0.95	0.344	−.0408535	.1170327
nch04	−.2100883	.0823633	−2.55	0.011	−.3715174	−.0486592
nch511	−.1190587	.0696115	−1.71	0.087	−.2554948	.0173774
nch1218	−.0001332	.066743	−0.00	0.998	−.130947	.1306807
age	.0686032	.0639068	1.07	0.283	−.0566519	.1938583
age2	−.0158264	.1301669	−0.12	0.903	−.2709489	.239296
age3	.133078	.0814011	1.63	0.102	−.0264653	.2926213
prof	−.6692923	.2205286	−3.03	0.002	−1.101521	−.2370641
mantech	−.6218779	.1017862	−6.11	0.000	−.8213753	−.4223806
skillmn	−.7479051	.1067488	−7.01	0.000	−.9571289	−.5386813
ptskill	−.5458932	.1018646	−5.36	0.000	−.7455442	−.3462423
unskill	−.2297057	.1566425	−1.47	0.143	−.5367193	.0773079
armed	−18.31647	8796.721	−0.00	0.998	−17259.57	17222.94
lninc	−.0566095	.0450624	−1.26	0.209	−.1449302	.0317111

• clogit $yvar $xvars if allwavesm==1, group(pid)

Table 9.10 Conditional logit model, balanced panel

Conditional (fixed-effects) logistic regression					Number of obs=	14736
					LR chi2 (17)=	555.50
					Prob > chi2=	0.0000
Log likelihood=−5180.5548					Pseudo R2=	0.0509
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
widowed	−.5365557	.1873604	−2.86	0.004	−.9037754	−.1693361
nvrmar	.061652	.1683354	0.37	0.714	−.2682792	.3915833
divsep	−.007591	.1607726	−0.05	0.962	−.3226996	.3075175
hhsize	.0034678	.0463808	0.07	0.940	−.087437	.0943726
nch04	−.1619611	.0938517	−1.73	0.084	−.3459071	.021985
nch511	−.1421056	.0793582	−1.79	0.073	−.2976447	.0134336
nch1218	−.0056385	.0769573	−0.07	0.942	−.156472	.145195
age	.087333	.0726924	1.20	0.230	−.0551414	.2298074
age2	−.0573966	.1484076	−0.39	0.699	−.3482701	.2334769
age3	.1521037	.0932538	1.63	0.103	−.0306704	.3348778
prof	−.6797852	.266153	−2.55	0.011	−1.201436	−.1581349
mantech	−.5574733	.1115359	−5.00	0.000	−.7760796	−.338867
skillmn	−.7819144	.1197024	−6.53	0.000	−1.016527	−.547302
ptskill	−.5387026	.1149615	−4.69	0.000	−.764023	−.3133822
unskill	−.1814188	.1703173	−1.07	0.287	−.5152344	.1523969
armed	−18.30142	8809.09	−0.00	0.998	−17283.8	17247.2
lninc	−.1107779	.0540475	−2.05	0.040	−.2167092	−.0048467

This estimator uses variation over time in the dependent and independent variables, so time-invariant variables like education are excluded. Notice that current income (lninc) remains statistically significant, at least for the balanced sample.

9.3 DYNAMIC MODELS

To model dynamics in self-reported health problems we use dynamic panel probit specifications on both the balanced and unbalanced samples. We include previous health problems in our empirical models in order to capture state dependence, and the model can be interpreted as a first-order Markov process. The latent variable specification of the model that we estimate can be written as:

$$y_{it}^{*} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \gamma y_{it-1} + \alpha_i + \varepsilon_{it} \; (i=1, \ldots, N;\; t=2, \ldots T_i)$$

x_{it} is the set of observed variables that may be associated with the health indicator. To capture state dependence, y_{it-1} is an indicator for the individual's health state in the

previous wave and /is the parameter to be estimated. a_i is an individual-specific and time-invariant random component. ε_{it} is a time and individual-specific error term that is assumed to be normally distributed and uncorrelated across individuals and waves and uncorrelated with a_t. ε_{it} is assumed to be strictly exogenous, that is, the x_{it} are uncorrelated with ε_{is} for all t and s. The model can be estimated using pooled or random effects specifications. As we do not have a natural scale for the latent variable the variance of the idiosyncratic error term is restricted to equal one.

Correlated effects and initial conditions

To allow for the possibility that the observed regressors may be correlated with the individual effect we parameterize the individual effect (Mundlak 1978; Chamberlain 1984; Wooldridge 2005). This allows for correlation between the individual effects and the means of the regressors. In addition, because we are estimating dynamic models, we need to take account of the problem of initial conditions. Heckman (1981) describes two assumptions that are typically invoked concerning a discrete time stochastic process with binary outcomes. The same issues arise with an ordered categorical variable. The first assumption is that the initial observations are exogenous variables. This is invalid when the error process is not serially independent and the first observation is not the true initial outcome of the process. In our case, the latter condition is violated, while the former is unlikely to be correct. Treating the lagged dependent variable as exogenous when these assumptions are incorrect leads to inconsistent estimators. The second assumption is that the process is in equilibrium such that the marginal probabilities have approached their limiting values and can therefore be assumed to be time-invariant. This assumption is untenable when non-stationary variables such as age and time trends are included in the model, as here.

Wooldridge (2005) has suggested a convenient approach to deal with the initial-conditions problem in non-linear dynamic random effects models by modelling the distribution of the unobserved effect conditional on the initial value and any exogenous explanatory variables. This conditional maximum likelihood (CML) approach results in a likelihood function based on the joint distribution of the observations conditional on the initial observations. Parameterizing the distribution of the unobserved effects leads to a likelihood function that is easily maximized using pre-programmed commands with standard software (e.g. Stata). However it should be noted that the CML approach does specify a complete model for the unobserved effects and may therefore be sensitive to mis-specification.

We implement this approach by parameterizing the distribution of the individual effects as:

$$\alpha_i = \alpha_0 + \alpha_1 y_{i1} + \alpha_2 \bar{x}_i + u_i$$

where $\bar{x}_i$ is the average over the sample period of the observations on the exogenous variables. u_i is assumed to be distributed $\mathrm{N}(0,\sigma_u^2)$ and independent of the x variables, the initial conditions, and the idiosyncratic error term (ε_{it}). Substitution gives a model that has a random effects structure, with the regressors at time t augmented to include the

initial value y_{i1} and $\bar{x}_i$. Three features should be noted. First, this specification implies that the identified coefficients of any time-invariant regressors are composite effects of the relevant elements of β and α_2. Second, all time dummies must be dropped from $\bar{x}_i$ to avoid perfect colinearity. Third, the estimates of α_1 are of direct interest as they are informative about the relationship between the individual effect and initial health problems.

A new global is required for the list of regressors in the dynamic specification that includes lagged health problems (hprobt_1):

• global xvard "hprobt_1 male widowed nvrmar divsep deghdeg hndalev ocse hhsize nch04 nch511 nch1218 age age2 age3 nonwhite prof mantech skillmn ptskill unskill armed lninc"

This is augmented by the initial value of health problems and the withinmeans of the regressors in the Mundlak-Wooldridge specification:

• global xvarw "hprobt_1 male widowed nvrmar divsep deghdeg hndalev ocse hhsize nch04 nch511 nch1218 age age2 age3 nonwhite lninc prof mantech skillmn ptskill unskill armed hprobt1 mlninc"

The dynamic versions of the pooled probit models can be compared with and without the correlated effects specifications (see Tables 9.11 to 9.14, over page).

• dprobit $yvar $xvard, robust cluster (pid)

Table 9.11 Dynamic pooled probit model, unbalanced panel

Probit regression, reporting marginal effects — Number of obs= 52904
Wald chi2 (23)= 8676.04
Prob>chi2= 0.0000
Log pseudolikelihood=−14737.264 — Pseudo R2= 0.3666
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
hprobt_1*	.5513682	.0083611	73.13	0.000	.148968	.534981	.567756
male*	.0054383	.0034398	1.58	0.113	.458831	−.001304	.01218
widowed*	−.0054188	.0053879	−0.99	0.322	.090995	−.015979	.005141
nvrmar*	.0123055	.0056494	2.24	0.025	.151255	.001233	.023378
divsep*	.0173168	.0065664	2.77	0.006	.070505	.004447	.030187
deghdeg*	−.0315484	.0054942	−5.16	0.000	.110615	−.042317	−.02078
hndalev*	−.0254161	.0044029	−5.45	0.000	.217961	−.034046	−.016787
ocse*	−.0290528	.0040303	−6.86	0.000	.280584	−.036952	−.021154

hhsize	−.0010478	.0019797	−0.53	0.597	2.77162	−.004928	.002832
nch04	−.0131115	.0048005	−2.73	0.006	.14205	−.02252	−.003703
nch511	−.0040938	.0033963	−1.21	0.228	.262835	−.01075	.002563
nch1218	−.0025836	.0039299	−0.66	0.511	.174675	−.010286	.005119
age	.0134217	.0021943	6.10	0.000	47.5315	.009121	.017722
age2	−.0232394	.0042471	−5.46	0.000	25.6937	−.031564	−.014915
age3	.0138831	.0025619	5.41	0.000	15.3677	.008862	.018904
nonwhite*	.0423465	.0090394	5.22	0.000	.04166	.02463	.060063
prof*	−.0542639	.0065213	−6.22	0.000	.034648	−.067045	−.041482
mantech*	−.0543756	.0038217	−12.15	0.000	.188568	−.061866	−.046885
skillmn*	−.0638684	.0038299	−12.82	0.000	.120879	−.071375	−.056362
ptskill*	−.0479479	.0042243	−9.30	0.000	.084871	−.056227	−.039668
unskill*	−.047715	.006502	−5.76	0.000	.025877	−.060459	−.034971
armed*	−.0873949	.0157578	−2.02	0.043	.000567	−.11828	−.05651
lninc	−.0197587	.0025464	−7.74	0.000	9.50969	−.024749	−.014768
obs. P	.1600635						
pred. P	.1002726 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P> | z | correspond to the test of the underlying coefficient being 0

• dprobit $yvar $xvard if allwavesm==1, robust cluster(pid)

Table 9.12 Dynamic pooled probit model, balanced panel

Probit regression, reporting marginal effects — Number of obs= 42475
Wald chi2 (22)= 6363.49
Prob>chi2= 0.0000
Log pseudolikelihood=−11386.258 — Pseudo R2= 0.3635
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
hprobt_1*	.5567582	.009626	64.54	0.000	.139753	.537892	.575625
male*	.0034882	.0037291	0.94	0.349	.449323	−.003821	.010797
widowed*	.0061219	.0059349	−1.01	0.312	.08259	−.017754	.00551
nvrmar*	.0115572	.0061944	1.92	0.054	.138905	−.000584	.023698
divsep*	.0105798	.0069309	1.58	0.114	.069264	−.003004	.024164
deghdeg*	−.0284846	.005955	−4.30	0.000	.114373	−.040156	−.016813
hndalev*	−.0197352	.0047126	−4.00	0.000	.225309	−.028972	−.010499
ocse*	−.0246546	.0043434	−5.42	0.000	.286733	−.033167	−.016142
hhsize	−.0023726	.002185	−1.09	0.277	2.79362	−.006655	.00191
nch04	−.0164735	.0051648	−3.19	0.001	.145945	−.026596	−.006351

nch511	−.0086144	.0037035	−2.33	0.020	.271148	−.015873	−.001356
nch1218	−.0028064	.0041305	−0.68	0.497	.176645	−.010902	.005289
age	.0132912	.0024651	5.38	0.000	47.3712	.00846	.018123
age2	−.0235046	.0048249	−4.86	0.000	25.3279	.032961	.014048
age3	.0138218	.0029488	4.68	0.000	14.913	.008042	.019601
nonwhite*	.0398803	.0111775	4.01	0.000	.032348	.017973	.061788
prof*	−.0563686	.0062513	−6.33	0.000	.035244	−.068621	−.044116
mantech*	−.0506674	.0040028	−10.83	0.000	.19475	−.058513	−.042822
skillmn*	−.0578174	.0040556	−11.03	0.000	.120706	−.065766	−.049869
ptskill*	−.0465828	.0043304	−8.65	0.000	.084473	−.05507	−.038095
unskill*	−.0398484	.0069763	−4.62	0.000	.026863	−.053522	−.026175
lninc	−.0235732	.0027778	−8.47	0.000	9.53283	−.029018	−.018129
obs. P	.1490995						
pred. P	.0920086 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient

• dprobit $yvar $xvarw, robust cluster (pid)

Table 9.13 Dynamic pooled probit model with initial conditions, unbalanced panel

Probit regression, reporting marginal effects — Number of obs= 52873
Wald chi2 (25)= 9646.12
Prob>chi2= 0.0000
Log pseudolikelihood=−14215.85 — Pseudo R2= 0.3886

(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
hprobt 1*	.4346145	.0097204	56.77	0.000	.148961	.415563	.453666
male*	.0065301	.0037233	1.76	0.079	.458665	−.000767	.013828
widowed*	−.0084472	.0057968	−1.42	0.156	.091048	−.019809	.002914
nvrmar*	.0112199	.0059519	1.94	0.053	.151306	−.000446	.022885
divsep*	.0124646	.0070027	1.85	0.065	.07049	−.00126	.02619
deghdeg*	−.0201285	.0064674	−2.92	0.004	.11068	−.032804	−.007453
hndalev*	−.0177621	.0049905	−3.42	0.001	.217824	−.027543	−.007981
ocse*	−.0202219	.0044999	−4.34	0.000	.280748	−.029041	−.011402
hhsize	−.0007849	.0020864	−0.38	0.707	2.77126	−.004874	.003304
nch04	−.0120821	.00501	−2.41	0.016	.142095	−.021902	−.002263
nch511	−.0049352	.0036505	−1.35	0.176	.262648	−.01209	.00222
nch1218	−.0024077	.0041306	−0.58	0.560	.17455	−.010504	.005688
age	.014693	.002402	6.10	0.000	47.5322	.009985	.019401

age2	−.0256332	.0046524	−5.50	0.000	25.6957	−.034752	−.016515
age3	.0151612	.0028102	5.39	0.000	15.3702	.009653	.020669
nonwhite*	.0291189	.0098246	3.22	0.001	.041685	.009863	.048375
lninc	−.0007486	.0036185	−0.21	0.836	9.50963	−.007841	.006344
prof*	−.0520013	.0069322	−5.67	0.000	.03463	−.065588	−.038414
mantech*	−.048754	.004143	−10.23	0.000	.188508	−.056874	−.040634
ski llmn*	−.059874	.0040924	−11.45	0.000	.12078	−.067895	−.051853
ptskill*	−.0427041	.0045889	−7.80	0.000	.084902	−.051698	−.03371
unskill*	−.0421848	.0067476	−5.08	0.000	.025892	−.05541	−.02896
armed*	−.0845266	.0147811	−2.21	0.027	.000567	−.113497	−.055556
hprobt1*	.1745513	.0084355	26.43	0.000	.12148	.158018	.191085
mlninc	−.0309038	.0052746	−5.85	0.000	9.50835	−.041242	−.020566
obs. P	.1600628						
pred. P	.0984209 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient being 0

• dprobit $yvar $xvarw if allwavesm==1, robust cluster (pid)

Table 9.14 Dynamic pooled probit model with initial conditions, balanced panel

Probit regression, reporting marginal effects — Number of obs= 42475
Wald chi2 (24)= 7218.66
Prob>chi2= 0.0000
Log pseudolikelihood=−10971.013 — Pseudo R2= 0.3867

(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar	[95% C.I.]	
hprobt_1*	.4433913	.011059	51.58	0.000	.139753	.421716	.465066
male*	.0053184	.0040284	1.32	0.186	.449323	−.002577	.013214
widowed*	−.0090136	.0063071	−1.38	0.166	.08259	−.021375	.003348
nvrmar*	.0104135	.0065046	1.65	0.100	.138905	−.002335	.023162
divsep*	.0051891	.0073157	0.72	0.470	.069264	−.009149	.019528
deghdeg*	−.0164034	.0070901	−2.18	0.029	.114373	−.0303	−.002507
hndalev*	−.0118831	.0053948	−2.14	0.032	.225309	−.022457	.00131
ocse*	−.0157712	.0048489	−3.16	0.002	.286733	−.025275	−.006267
hhsize	−.0025417	.0022953	−1.11	0.268	2.79362	−.00704	.001957
nch04	−.0150942	.0053878	−2.80	0.005	.145945	−.025654	−.004534
nch511	−.0085618	.0039585	−2.16	0.030	.271148	−.01632	−.000803
nch1218	−.0011601	.0043455	−0.27	0.790	.176645	−.009677	.007357
age	.0144276	.0027075	5.32	0.000	47.3712	.009121	.019734

age2	−.0258634	.0052879	−4.88	0.000	25.3279	−.036227	−.015499
age3	.0152666	.0032255	4.73	0.000	14.913	.008945	.021589
nonwhite*	.0240862	.0121326	2.15	0.032	.032348	.000307	.047866
lninc	−.0043908	.0039072	−1.12	0.261	9.53283	−.012049	.003267
prof*	−.0542811	.0065987	−5.83	0.000	.035244	−.067214	−.041348
mantech*	−.0449159	.0043334	−9.03	0.000	.19475	−.053409	−.036423
ski llmn*	−.0539528	.0043165	−9.84	0.000	.120706	−.062413	−.045493
ptskill*	−.0416337	.0047419	−7.23	0.000	.084473	−.050928	−.03234
unskill*	−.0344624	.0072099	−4.00	0.000	.026863	−.048594	−.020331
hprobt1*	.1724311	.0095016	23.62	0.000	.112066	.153808	.191054
mlninc	−.0304624	.0057946	−5.25	0.000	9.5323	−.04182	−.019105
obs. P	.1490995						
pred. P	.0900295 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and P>|z| correspond to the test of the underlying coefficient being 0

There is some reduction in the partial effect of lagged health (hprob_1) when the adjustment for initial conditions is included: from around 0.55 to around 0.44. But the state dependence effect remains large.

Similarly, the random effects specifications can also be extended to include dynamics (see Tables 9.15 to 9.18, over page).

• xtprobit $yvar $xvard, intp(24)

Table 9.15 Dynamic random effects probit model, unbalanced panel

Random-effects probit regression		Number of obs=	48560
Group variable (i):pid		Number of groups=	6070
Random effects u_i ~ Gaussian		Obs per group: min=	8
		avg=	8.0
		max=	8
		Wald chi2 (23)=	1088.01
Log likelihood=−12 619.818		Prob>chi2=	0.0000

hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	−.1192647	.0599106	−1.99	0.047	−.2366872	−.0018421
widowed	−.3011138	.0808099	−3.73	0.000	−.4594983	−.1427294
nvrmar	.1850041	.0719871	2.57	0.010	.043912	.3260961
divsep	.0914589	.0739631	1.24	0.216	−.0535061	.2364239
deghdeg	−.3919638	.1174985	−3.34	0.001	−.6222566	−.1616711
hndalev	−.2111355	.0867028	−2.44	0.015	−.3810699	−.0412012
ocse	−.3913465	.0783682	−4.99	0.000	−.5449454	−.2377476

hhsize	−.0195375	.0221159	−0.88	0.377	−.0628839	.0238089
nch04	−.1551483	.0468387	−3.31	0.001	−.2469505	−.0633462
nch511	−.1323934	.0374751	?3.53	0.000	−.2058433	−.0589435
nch1218	−.0285701	.0382921	−0.75	0.456	−.1036214	.0464811
age	.159061	.0292788	5.43	0.000	.1016757	.2164463
age2	−.3008401	.0583309	−5.16	0.000	−.4151666	−.1865135
age3	.208999	.0364707	5.73	0.000	.1375178	.2804803
nonwhite	.7945646	.1517943	5.23	0.000	.4970532	1.092076
lninc	−.0458109	.0299387	−1.53	0.126	−.1044898	.0128679
prof	−.5693845	.1205951	−4.72	0.000	−.8057465	−.3330226
mantech	−.4969237	.0548426	−9.06	0.000	−.6044133	−.3894341
skillmn	−.6597481	.0603996	−10.92	0.000	−.7781291	−.5413671
ptskill	−.4792066	.0596291	−8.04	0.000	−.5960774	−.3623357
unskill	−.3314323	.0901974	−3.67	0.000	−.508216	−.1546486
armed	−6.478364	5441.711	−0.00	0.999	−10672.04	10659.08
mlninc	−.6156707	.0696439	−8.84	0.000	−.7521702	−.4791711
_cons	1.591711	.7064546	2.25	0.024	.2070855	2.976337
/lnsig2u	1.069177	.0425637			.9857541	1.152601
sigma_u	1.706746	.0363227			1.637019	1.779443
rho	.7444405	.0080977			.7282485	.7599856

Likelihood-ratio test of rho=0: chibar2 (01)=1.0e+04 Prob >=chibar2= 0.000

• xtprobit $yvar $xvard if allwavesm==1, intp (24)

Table 9.16 Dynamic pooled probit model, balanced panel

Random-effects probit regression		Number of obs=	42490
Group variable (i): pid		Number of groups=	6070
Random effects u i ~ Gaussian		Obs per group:min=	7
		avg=	7.0
		max=	7
		Wald chi2 (23)=	2953.42
Log likelihood=−11040.906		Prob>chi2=	0.0000

hprob	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
hprobt_1	1.082657	.0320234	33.81	0.000	1.019892	1.145421
male	−.0220624	.03726	−0.59	0.554	−.0950907	.0509659
widowed	−.1334305	.0632548	−2.11	0.035	−.2574076	−.0094534
nvrmar	.1193202	.0566	2.11	0.035	.0083862	.2302542
divsep	.1242645	.0604756	2.05	0.040	.0057345	.2427944

deghdeg	−.4053264	.0724259	−5.60	0.000	−.5472786	−.2633741
hndalev	−.2729744	.052838	−5.17	0.000	−.376535	−.1694138
ocse	−.3283334	.0478804	−6.86	0.000	−.4221774	−.2344895
hhsize	−.0265543	.019319	−1.37	0.169	−.064419	.0113103
nch04	−.1068818	.0427582	−2.50	0.012	−.1906864	−.0230772
nch511	−.0800314	.0325579	−2.46	0.014	−.1438436	−.0162191
nch1218	−.0067642	.0366142	−0.18	0.853	−.0785267	.0649983
age	.1124985	.023738	4.74	0.000	.0659729	.1590241
age2	−.2072082	.0470305	−4.41	0.000	−.2993863	−.11503
age3	.1329085	.0292072	4.55	0.000	.0756635	.1901535
nonwhite	.4526229	.0939747	4.82	0.000	.268436	.6368099
prof	−.6068532	.1029152	−5.90	0.000	−.8085634	−.4051431
mantech	−.4817259	.0478139	−10.08	0.000	−.5754393	−.3880124
skillmn	−.5917575	.0544084	−10.88	0.000	−.6983961	−.4851189
ptskill	−.4552731	.0550294	−8.27	0.000	−.5631288	−.3474174
unskill	−.3436122	.0841974	−4.08	0.000	−.5086361	−.1785882
armed	−7.386609	12231.98	−0.00	1.000	−23981.62	23966.85
lninc	−.1763811	.0244266	−7.22	0.000	−.2242564	−.1285059
_cons	−1.734423	.4179103	−4.15	0.000	−2.553512	−.9153335
/lnsig2u	− .147883	.048054			−.2420671	−.0536988
sigma_u	.928726	.0223145			.8860042	.9735078
rho	.4630965	.0119481			.439777	.4865785

Likelihood-ratio test of rho=0: chibar2 (01)=690.70 Prob>=chibar2= 0.000

• xtprobit $yvar $xvarw, intp(24)

Table 9.17 Dynamic pooled probit model with initial conditions, unbalanced panel

Random-effects probit regression		Number of obs=	52873
Group variable (i) : pid		Number of groups=	9206
Random effects u_i ~ Gaussian		Obs per group: min=	1
		avg=	5.7
		max=	7
		Wald chi2 (25)=	5702.41
Log likelihood=−13567.655		Prob > chi2=	0.0000

hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hprobt_1	.7678608	.0281689	27.26	0.000	.7126509	.8230708
male	.029484	.0319731	0.92	0.356	−.0331822	.0921502
widowed	−.1197592	.0536579	−2.23	0.026	−.2249267	−.0145916
nvrmar	.0777616	.0489031	1.59	0.112	−.0180867	.1736099

divsep	.1034386	.0529565	1.95	0.051	−.0003543	.2072314
deghdeg	−.1808878	.0647961	−2.79	0.005	−.3078859	−.0538897
hndalev	−.1598575	.0470894	−3.39	0.001	−.252151	−.0675641
ocse	−.1873981	.0419127	−4.47	0.000	−.2695455	−.1052507
hhsize	−.0089249	.0167439	−0.53	0.594	−.0417425	.0238926
nch04	−.0855237	.0379835	−2.25	0.024	−.1599699	−.0110774
nch511	−.0486441	.0285896	−1.70	0.089	−.1046788	.0073905
nch1218	−.0035524	.0328512	−0.11	0.914	−.0679396	.0608348
age	.1124817	.019857	5.66	0.000	.0735627	.1514008
age2	−.2008239	.038896	−5.16	0.000	−.2770587	−.1245891
age3	.1262069	.0238167	5.30	0.000	.079527	.1728868
nonwhite	.2837638	.0699037	4.06	0.000	.146755	.4207726
lninc	−.0170018	.0265107	−0.64	0.521	−.0689618	.0349581
prof	−.5034843	.090729	−5.55	0.000	−.6813098	−.3256587
mantech	−.4088152	.043465	−9.41	0.000	−.4940051	−.3236254
skillmn	−.5454347	.048801	−11.18	0.000	−.6410829	−.4497865
ptskill	−.375294	.0485184	−7.74	0.000	−.4703882	−.2801997
unskill	−.3190351	.0780555	−4.09	0.000	−.4720211	−.1660491
armed	−1.092631	.8306451	−1.32	0.188	−2.720665	.5354038
hprobt1	1.672542	.0475462	35.18	0.000	1.579353	1.765731
mlninc	−.2788631	.0426786	−6.53	0.000	−.3625115	−.1952146
_cons	−1.002668	.4115023	−2.44	0.015	−1.809197	−.1961379
/lnsig2u	−.2094432	.0409362			−.2896766	−.1292098
sigma_u	.9005752	.018433			.8651622	.9374378
rho	.4478298	.0101226			.428083	.4677424

Likelihood-ratio test of rho=0: chibar2 (01)=1296.39 Prob>=chibar2= 0.000

• xtprobit $yvar $xvarw if allwavesm==1, intp(24)

Table 9.18 Dynamic pooled probit model with initial conditions, balanced panel

Random-effects probit regression		Number of obs=	42490
Group variable (i): pid		Number of groups=	6070
Random effects u_i ~ Gaussian		Obs per group: min=	7
		avg=	7.0
		max=	7
		Wald chi2 (25)=	3975.35
Log likelihood=−10429.3		Prob>chi2=	0.0000

hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hprobt 1	.7831548	.0316115	24.77	0.000	.7211974	.8451121
male	.0136385	.0378468	0.36	0.719	−.0605399	.087817
widowed	−.1715777	.0647571	−2.65	0.008	−.2984993	−.044656
nvrmar	.0957191	.0573502	1.67	0.095	−.0166851	.2081234
divsep	.0643965	.0618696	1.04	0.298	−.0568656	.1856587
deghdeg	−.159993	.0749971	−2.13	0.033	−.3069845	−.0130014
hndalev	−.1178776	.0546888	−2.16	0.031	−.2250658	−.0106894
ocse	−.1650021	.0491165	−3.36	0.001	−.2612687	−.0687355
hhsize	−.0268536	.019731	−1.36	0.174	−.0655256	.0118184
nch04	−.102412	.0434186	−2.36	0.018	−.1875108	−.0173131
nch511	−.0866443	.0334595	−2.59	0.010	−.1522237	−.0210649
nch1218	.0015585	.037366	0.04	0.967	−.0716775	.0747945
age	.129491	.0242341	5.34	0.000	.0819929	.176989
age2	−.2440243	.0479655	−5.09	0.000	−.338035	−.1500135
age3	.1551966	.02974	5.22	0.000	.0969072	.2134859
nonwhite	.248272	.095243	2.61	0.009	.0615991	.4349449
lninc	−.0463832	.0311862	−1.49	0.137	−.1075071	.0147407
prof	−.5516054	.1060574	−5.20	0.000	−.759474	−.3437367
mantech	−.3947172	.0488257	−8.08	0.000	−.4904139	−.2990205
skillmn	−.519226	.0550982	−9.42	0.000	−.6272164	−.4112356
ptskill	−.384451	.0554431	−6.93	0.000	−.4931176	−.2757845
unskill	−.2666127	.0857013	−3.11	0.002	−.4345843	−.0986412
armed	−6.97209	13733.1	−0.00	1.000	−26923.35	26909.4
hprobt1	1.726174	.0565217	30.54	0.000	1.615394	1.836955
mlninc	−.3129958	.0512549	−6.11	0.000	−.4134536	−.212538
_cons	−.5294052	.4998143	−1.06	0.290	−1.509023	.4502127
/lnsig2u	−.1809886	.0464303			−.2719904	−.0899869
sigma_u	.9134795	.0212066			.8728468	.9560038
rho	.454876	.011513			.4324185	.4775184

Likelihood-ratio test of rho=0: chibar2 (01)=1083.43 Prob>=chibar2= 0.000

The Heckman estimator

The Wooldridge approach is attractive for its simplicity, but an alternative strategy for dealing with the initial-conditions problem, that makes weaker assumptions, is the estimator proposed by Heckman (1981). This has been implemented as a Stata command by Stewart (2006). The Heckman estimator specifies a reduced form for the latent variable in the initial wave which includes a vector of exogenous variables (z_{il}):

$$y^*_{il} = z_{il}\pi + \theta\alpha_i + \varepsilon_{il}$$

These exogenous variables should include some instruments that do not appear in the main equation. This is problematic for our empirical application.

Then the log-likelihood can be written in terms of the joint probability of the observed sequence of 1s and 0s, with a_t integrated out:

$$lnL = \sum_{i=1}^{n}\left\{ln\int_{-\infty}^{+\infty}\Phi[d_{i1}(z_{i1}\pi+\theta\alpha)]\prod_{t=2}^{T}(\Phi[d_{it}(x_{it}\beta+\gamma y_{it-1}+\alpha)])f(\alpha)d\alpha\right\}$$

Like the conventional random effects probit model, this can be estimated by Gauss-Hermite quadrature. First a global has to be specified for the list of instruments used to predict the initial value. Then the model is estimated using the command redprob, which specifies the individual (i) and time indices (t) as subcommands. The default starting values for the maximum likelihood estimation are taken from a separate probit for the reduced form and a pooled probit for the remaining periods:

• global z0 “malet1 widowedt1 nvrmart1 divsept1 deghdegt1 hndalevt1 ocset1 hhsizet1 nch04t1 nch511t1 nch1218t1 aget1 age2t1 age3t1 nonwhitet1 proft1 mantecht1 skillmnt1 ptskillt1 unskillt1 armedt1 lninct1”

• redprob $yvar $xvard ($z0), i (pid) t(wavenum), quadrat (24)

First the command reports the starting values:

Pooled Probit Model for t>1
Iteration 0 log likelihood = −23265.866
Iteration 1 log likelihood = −14962.392
Iteration 2 log likelihood = −14740.692
Iteration 3 log likelihood = −14737.266
Iteration 4 log likelihood = −14737.264
Iteration 5 log likelihood = −14737.264

Probit regression	Number of obs = 52904
	LRchi2 (23) = 17057.20
	Prob > chi2 = 0.0000

Log likelihood = −14737 264					Pseudo R2 =	0.3666
hprob	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
hprobt_1	1.836876	.0182638	100.57	0.000	1.80108	1.872673
male	.0308751	.0174072	1.77	0.076	−.0032424	.0649926
widowed	−.0313267	.0311382	−1.01	0.314	−.0923565	.0297032
nvrmar	.0679004	.0292035	2.33	0.020	.0106625	.1251382
divsep	.0935945	.0322579	2.90	0.004	.0303702	.1568188
deghdeg	−.1983709	.0355797	−5.58	0.000	−.2681058	−.128636
hndalev	−.1527605	.0251221	−6.08	0.000	−.2019989	−.1035221
ocse	−.1734283	.0223108	−7.77	0.000	−.2171566	−.1296999
hhsize	−.0059588	.0109008	−0.55	0.585	−.027324	.0154065
nch04	−.0745616	.0266679	−2.80	0.005	−.1268297	−.0222935
nch511	−.0232807	.0184472	−1.26	0.207	−.0594365	.0128752
nch1218	−.014692	.0225154	−0.65	0.514	−.0588214	.0294374
age	.0763259	.011709	6.52	0.000	.0533767	.099275
age2	−.1321569	.0228205	−5.79	0.000	−.1768843	−.0874294
age3	.0789496	.0139107	5.68	0.000	.0516852	.106214
nonwhite	.2133412	.0382898	5.57	0.000	.1382945	.2883879
prof	−.3939215	.0601556	−6.55	0.000	−.5118243	−.2760188
mantech	−.3569573	.0286281	−12.47	0.000	−.4130673	−.3008473
skillmn	−.4555487	.033263	−13.70	0.000	−.520743	−.3903544
ptskill	−.3257354	.0339987	−9.58	0.000	−.3923716	−.2590992
unskill	−.3350907	.0558528	−6.00	0.000	−.4445601	−.2256212
armed	−.9475817	.5916095	−1.60	0.109	−2.107115	.2119515
lninc	−.1123626	.0141489	−7.94	0.000	−.1400938	−.0846313
_cons	−1.655159	.2164497	−7.65	0.000	−2.079392	−1.230925

Probit Model for t=1

Iteration 0: log likelihood=−4119.4516

Iteration 1: log likelihood=−3570.1143

Iteration 2: log likelihood=−3554.2801

Iteration 3: log likelihood=−3554.19

Iteration 4: log likelihood=−3554.19

Probit regression					Number of obs=	10247
					LR chi2 (22)=	1130.52
					Prob>chi2=	0.0000
Log likelihood=−3554.19					Pseudo R2=	0.1372
hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
malet1	−.0068734	.0362775	−0.19	0.850	−.0779759	.0642292
widowedt1	−.0034003	.0625214	−0.05	0.957	−.1259399	.1191393

nvrmart1	.0557137	.0584071	0.95	0.340	−.0587621	.1701896
divsept1	.1990238	.0686795	2.90	0.004	.0644144	.3336332
deghdegt1	−.3331031	.0800086	−4.16	0.000	−.489917	−.1762892
hndalevt1	−.1806528	.0531712	−3.40	0.001	−.2848665	−.0764391
ocset1	−.3234213	.0477855	−6.77	0.000	−.417079	−.2297635
hhsizet1	.0094386	.0212157	0.44	0.656	−.0321434	.0510205
nch04t1	−.1553132	.052663	−2.95	0.003	−.2585308	−.0520956
nch511t1	−.05919	.0389182	−1.52	0.128	−.1354683	.0170882
nch1218t1	−.0761862	.0412182	−1.85	0.065	−.1569723	.0045999
aget1	.0476419	.0203852	2.34	0.019	.0076875	.0875962
age2t1	−.0735117	.041254	−1.78	0.075	−.1543682	.0073447
age3t1	.0481789	.0259656	1.86	0.064	−.0027128	.0990705
nonwhitet1	.4265401	.0638679	6.68	0.000	.3013612	.551719
proft1	−.2650012	.119926	−2.21	0.027	−.5000519	−.0299506
mantecht1	−.4776225	.0633284	−7.54	0.000	−.6017438	−.3535012
skillmnt1	−.4761174	.0642507	−7.41	0.000	−.6020464	−.3501885
ptskillt1	−.5391528	.0733221	−7.35	0.000	−.6828615	−.3954441
unskillt1	−.5096275	.1156045	−4.41	0.000	−.7362082	−.2830467
armedt1	−.5259454	.4892789	−1.07	0.282	−1.484914	.4330238
lninct1	−.1461868	.0298004	−4.91	0.000	−.2045946	−.087779
_cons	−.5782361	.3985882	−1.45	0.147	−1.359455	.2029824

Then the final output (Table 9.19):

Table 9.19 Heckman estimator of dynamic random effects probit

Random-Effects Dynamic Probit Model — Number of obs= 64719

Wald chi2 (23)= 3202.16

Log likelihood=−15523.372 — Prob>chi2= 0.0000

hprob	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hprob						
hprobt_1	.8054048	.0302172	26.65	0.000	.7461802	.8646293
male	−.0029672	.0377062	−0.08	0.937	−.0768701	.0709357
widowed	−.0802493	.0574424	−1.40	0.162	−.1928344	.0323357
nvrmar	.0224521	.058573	0.38	0.701	−.0923488	.137253
divsep	.1607117	.0578991	2.78	0.006	.0472314	.2741919
deghdeg	−.4844789	.0771783	−6.28	0.000	−.6357456	−.3332123
hndalev	−.3364978	.0571368	−5.89	0.000	−.4484839	−.2245117
ocse	−.4261034	.0501261	−8.50	0.000	−.5243488	−.327858

hhsize	−.0066062	.0188406	−0.35	0.726	−.0435332	.0303207
nch04	−.1026608	.0419096	−2.45	0.014	−.1848022	−.0205194
nch511	−.0598588	.0323816	−1.85	0.065	−.1233255	.003608
nch1218	.0110948	.0353891	0.31	0.754	−.0582665	.0804562
age	.0913202	.0232513	3.93	0.000	.0457485	.1368919
age2	−.165139	.0450961	−3.66	0.000	−.2535256	−.0767523
age3	.1159508	.0273435	4.24	0.000	.0623584	.1695431
nonwhite	.5743562	.0826363	6.95	0.000	.4123921	.7363203
prof	−.6910387	.1115201	−6.20	0.000	−.909614	−.4724634
mantech	−.4921585	.0477005	−10.32	0.000	−.5856498	−.3986672
skillmn	−.6136016	.0526363	−11.66	0.000	−.7167669	−.5104363
ptskill	−.4320846	.0522031	−8.28	0.000	−.5344007	−.3297685
unskill	−.3424978	.0816927	−4.19	0.000	−.5026126	−.182383
armed	−1.041123	.8200141	−1.27	0.204	−2.648321	.566075
lninc	−.1353205	.0231107	−5.86	0.000	−.1806167	−.0900243
_cons	−1.978318	.4136655	−4.78	0.000	−2.789088	−1.167549
rfper1						
malet1	−.0706227	.0591113	−1.19	0.232	−.1864787	.0452332
widowedt1	−.0382902	.0954868	−0.40	0.688	−.2254408	.1488604
nvrmart1	.0062672	.0914728	0.07	0.945	−.1730161	.1855506
divsept1	.181747	.1078356	1.69	0.092	−.0296069	.3931008
deghdegt1	−.5763432	.130813	−4.41	0.000	−.832732	−.3199544
hndalevt1	−.298313	.0878715	−3.39	0.001	−.4705381	−.126088
ocset1	−.5915628	.0803316	−7.36	0.000	−.7490098	−.4341158
hhsizet1	.0144338	.0327746	0.44	0.660	−.0498032	.0786707
nch04t1	−.2593662	.0845408	−3.07	0.002	−.4250632	−.0936692
nch511t1	−.1016244	.061691	−1.65	0.099	−.2225365	.0192877
nch1218t1	−.0785009	.0626013	−1.25	0.210	−.2011973	.0441954
aget1	.0221281	.0322885	0.69	0.493	−.0411561	.0854123
age2t1	−.0383784	.0654221	−0.59	0.557	−.1666033	.0898465
age3t1	.0453517	.0411373	1.10	0.270	−.035276	.1259793
nonwhitet1	.7732582	.108174	7.15	0.000	.5612411	.9852754
proft1	−.467453	.2009746	−2.33	0.020	−.861356	−.07355
mantecht1	−.5984927	.0969756	−6.17	0.000	−.7885613	−.4084241

hprob	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
skillmnt1	−.6358015	.0998663	−6.37	0.000	−.8315358	−.4400671
ptskillt1	−.602463	.1084195	−5.56	0.000	−.8149613	−.3899646
unskillt1	−.5782615	.1708122	−3.39	0.001	−.9130471	−.2434758
armedt1	−.8677926	.8062199	−1.08	0.282	−2.447954	.7123693
lninct1	−.0975383	.0453351	−2.15	0.031	−.1863934	−.0086833

_cons	−1.144854	.6178439	−1.85	0.064	−2.355806	.0660982
/logitrho	.1840367	.0444404	4.14	0.000	.0969352	.2711382
/ltheta	.0777463	.0464464	1.67	0.094	−.013287	.1687797
rho	.5458798	.0110165	49.55	0.000	.5242148	.5673723
theta	1.080848	.0502016	21.53	0.000	.9868008	1.183859

LR test of rho=0: chi2(1)=5536.17
Prob>chi2=0.0000

10
Non-response and attrition bias

10.1 INTRODUCTION

The objective of this chapter is to explore the existence of health-related non-response in panel data and its consequences for modelling the association between socioeconomic status (SES) and health problems. It builds on the results of the previous chapter and on a paper by Jones, Koolman and Rice (2006) that analyses self-assessed health rather than health problems.

Using panel data, such as the British Household Panel Survey (BHPS), to analyse longitudinal models of health problems creates a risk that the results will be contaminated by bias associated with longitudinal non-response. There are drop-outs from the panels at each wave and some of these may be related directly to health: owing to deaths, serious illness and people moving into institutional care. In addition, other sources of non-response may be indirectly related to health, for example divorce may increase the risk of non-response and also be associated with poorer health than average. The long-term survivors who remain in the panel are likely to be healthier on average compared with the sample at wave 1. The health of survivors will tend to be higher than that of the population as a whole and their rate of decline in health will tend to be lower. Also, the socioeconomic status of the survivors may not be representative of the original population who were sampled at wave 1. Failing to account for non-response may result in misleading estimates of the relationship between health and socioeconomic characteristics.

The pattern of non-response can be tabulated to show how the sample size and composition evolve across the eight waves of the BHPS. The data used to construct the table include the number of observations that are available at each wave and the corresponding number of drop-outs and re-joiners between waves. These are expressed as wave-on-wave survival and drop-out rates. The survival rate is the percentage of original sample members remaining at wave *t*. The drop-out rate is the percentage of the number of drop-outs between waves *t-1* and *t* to the number of observations at *t-1*. The raw drop-out rate excludes re-joiners, while the net drop-out rate includes them. These measures are constructed from the indicator of non-response insampm. Here the variable is recoded to system missing (.) for the non-responders:

```
• gen miss=insampm
  • replace miss=.if insampm==0
```

Then the following program calculates the statistics that are needed, looping through the waves of the panel (wavenum):

```
• program define table
        {
            quietly summ miss if wavenum==1
                scalar NO=r (N)
                forvalues j=2 (1) 8{
                    display “wavenum== ” ‘j’
                    quietly summ miss if (wavenum == ‘j’?1
                scalar N1=r (N)
                quietly summ miss if (wavenum== ‘j’ & miss
                [_n?1] ~= .)
                scalar N2=r (N)
                quietly summ miss if (wavenum== ‘j’ & miss
                [_n?1]==.)
                scalar N3=r(N)
                quietly summ miss if (wavenum==‘j’)
                scalar N4=r(N)
                scalar dropout=N1?N2
                scalar rejoiner=N3
                scalar rattr=( (N1?N2) /N1)
                scalar nattr=( (N1?N4) /N1)
                scalar surv=N4/N0
                display “No. individuals at wave=” ‘j’?1
                “=” N1
                display “No. individuals at wave=” ‘j’ “=” N4
                display “Survival rate= “surv ” Drop outs
                =” dropout “Re-joiners =” rejoiner
                display “Raw Attrition rate=“rattr” Net
                Attrition rate=“nattr
                display” ”
                    }
    }
    end
```

• table

Running the program produces the following output, which can be used to tabulate the number of individuals in the sample at each wave, the number of drop-outs, the number of re-joiners, the survival rate and the raw and net drop-out rates (see Table 1 in Jones, Koolman and Rice 2006):

```
wavenum==2
    No. individuals at wave=1=10247
    No. individuals at wave=2=8954
    Survival rate=.87381673 Drop outs=1410 Re-joiners=117
    Raw Attrition rate=.13760125 Net Attrition rate=.12618327
```

```
wavenum==3
    No. individuals at wave=2=8954
    No. individuals at wave=3=8024
    Survival rate=.78305846 Drop outs=1036 Re-joiners=106
    Raw Attrition rate=.11570248 Net Attrition rate=.10386419

wavenum==4
    No. individuals at wave=3=8024
    No. individuals at wave=4=7874
    Survival rate=.76842003 Drop outs=237 Re-joiners=87
    Raw Attrition rate=.02953639 Net Attrition rate=.01869392

wavenum==5
    No. individuals at wave=4=7874
    No. individuals at wave=5=7451
    Survival rate=.72713965 Drop outs=518 Re-joiners=95
    Raw Attrition rate=.06578613 Net Attrition rate=.05372111

wavenum==6
    No. individuals at wave=5=7451
    No. individuals at wave=6=7379
    Survival rate=.7201132 Drop outs=168 Re-joiners=96
    Raw Attrition rate=.02254731 Net Attrition rate=.00966313

wavenum==7
    No. individuals at wave=6=7379
    No. individuals at wave=7=7128
    Survival rate=.69561823 Drop outs=341 Re-joiners=90
    Raw Attrition rate=.04621222 Net Attrition rate=.03401545

wavenum==8
    No. individuals at wave=7=7128
    No. individuals at wave=8=6861
    Survival rate=.66956182 Drop outs=358 Re-joiners=91
    Raw Attrition rate=.05022447 Net Attrition rate=.03745791
```

Drop-out rates are highest between waves 1 and 2, with a raw attrition rate of 14%, and the rate tends to decline over time, with a rate of 5% between waves 7 and 8. By wave 8 the original sample of 10,247 has been reduced to 6,861.

Nicoletti and Peracchi (2005) provide a taxonomy of reasons for non-participation in surveys. Non-response can arise because of:

1 Demographic events such as death.
2 Movement out of the scope of the survey such as institutionalization or emigration.
3 Refusal to respond at subsequent waves.
4 Absence of the person at the address.
5 Other types of non-contact.

To these points, we would add item non-response for any of the variables used in the model of health problems, which eliminates these observations from the sample. The notion of attrition, commonly used in the survey methods literature, is usually restricted to points 3, 4 and 5. However our concern is with any longitudinal non-response that leads to missing observations in the panel data regression analysis. In fact it is points 1 and 2—death and incapacity—that are likely to be particularly relevant as sources of health-related non-response. The original sample consists of those who provide a full interview and usable information on health problems at the first wave of the BHPS. Non-response encompasses all of those who fail to provide usable observations for the model of health problems at subsequent waves.

We take a representative sample of individuals at wave 1 and follow them for the eight years of the BHPS sample used in our application. The sample of interest is those n original individuals observed over a full T-year period (T=8). A fully observed sample from this population would consist of nT observations. Owing to non-response we only observe $\sum_{i=1}^{n} T_i$ observations.

The reasons for having incomplete observations include attrition (as conventionally defined in the survey methods literature) as well as individuals becoming ineligible because of incapacity or death. This creates a problem of *incidental truncation:* we are interested in the association between health and SES for our n individuals over the full T waves. However, the frailer individuals are more likely to die or drop out before the end of the observation period, and their levels of health problems and SES are unobservable. This means that the remaining observed sample of survivors may contain fewer frail individuals—this is the source of potential bias in the relationship between health and SES across our sample of individuals.

10.2 TESTING FOR NON-RESPONSE BIAS

To provide an initial test for non-response bias we use the simple variable addition tests proposed by Verbeek and Nijman (1992; p. 688). These tests work by constructing variables that reflect the pattern of survey response provided by each individual respondent. Recall from Chapter 2 that we created indicators of whether an individual appears in the next wave (nextwavem) and whether they appear in the balanced panel (allwavesm), along with the number of waves that the individual is in, in the panel (Ti):

- sort pid wavenum
  - by pid: gen nextwavem=insampm [_n+1]
  - gen allwavesm=.
  - recede allwavesm.=0 if Ti~=8
  - recede allwavesm.=1 if Ti==8
  - gen numwavesm=.
  - replace numwavesm=Ti

Each of these three variables is an indicator of how the individual responds to the survey. There should be no intrinsic reason that survey response should have an effect on the individual's health. However if there is selection bias, such that those who do not respond have systematically different health from those who do, there will be a statistical association between the new variables and individuals' health. The tests work by adding the new variables to the pooled and random-effects probit models that are estimated with the unbalanced sample. The statistical significance of the added variables provides a test for non-response bias. This can be done for both static and dynamic specifications. The models are run quietly as we are only interested in the test statistics and test is used to compute a chi-squared test:

• *i) WITH Ti
  • quietly probit $yvar $xvard Ti, robust cluster (pid)
  • test Ti=0

(1) Ti=0
chi2 (1)=23.08
Prob > chi2−0.0000

• quietly xtprobit $yvar $xvars Ti, intp(24)
  • test Ti=0

(1) [hprob]Ti=0
chi2 (1)=1.95
Prob > chi2–0.1624

• * ii) WITH ALLWAVESM
  • quietly probit $yvar $xvars allwavesm, robust cluster(pid)

• test allwavesm=0

(1) allwavesm= 0
chi2(1)=15.12
Prob > chi2=0.0001

• quietly xtprobit $yvar $xvars allwavesm, intp(24)
  • test allwavesm=0

(1) [hprob]allwavesm=0
chi2 (1)=6.92
Prob > chi2=0.0085

• * ill) WITH Sit+1
  • quietly probit $yvar $xvars nextwavem, robust cluster(pid)
  • test nextwavem=0

(1) nextwavem= 0
chi2(1)=32.35
Prob > chi2=0.0000

• quietly xtprobit $yvar $xvars nextwavem, intp (24)
• test nextwavem=0

(1) [hprob]nextwavem=0
chi2(1)=30.84
Prob > chi2=0.0000

• * DYNAMIC MUNDLAK/WOOLDRIDGE VERSION
• * i) WITH Ti
• quietly probit $yvar $xvarw Ti, robust cluster (pid)
• test Ti=0

(1) Ti=0
chi2(1)=7.50
Prob > chi2=0.0062

• quietly xtprobit $yvar $xvarw Ti, intp (24)
• test Ti=0

(1) [hprob]Ti=0
chi2 (1)=3.58
Prob > chi2=0.0584

• * ii) WITH ALLWAVESM
• quietly probit $yvar $xvarw allwavesm, robust cluster(pid)
• test allwavesm=0

(1) allwavesm=0
chi2 (1)= 9.80
Prob > chi2= 0.0017

• quietly xtprobit $yvar $xvarw allwavesm, intp (24)
• test allwavesm=0

(1) [hprob]allwavesm=0
chi2 (1)= 6.60
Prob > chi2= 0.0102

• * iii) WITH Sit+1
• quietly probit $yvar $xvarw nextwavem, robust cluster(pid)
• test nextwavem=0

```
(1) nextwavem=0
          chi2 (1)=    16.01
       Prob > chi2=    0.0001
```

```
• quietly xtprobit $yvar $xvarw nextwavem, intp (24)
   • test nextwavem=0
```

```
(1) [hprob]nextwavem=0
          chi2(1)=    19.02
       Prob > chi2=    0.0000
```

With a couple of exceptions for the random-effects models, these tests reject the null hypothesis ($p<0.05$), suggesting that there is a problem of attrition bias.

The intuition behind these tests is that, if non-response is random, indicators of an individual's pattern of survey responses *(r)* should not be associated with the outcome of interest *(y)* after controlling for the observed covariates *(x)*: in other words, it tests a conditional independence condition:

$E(y|x,r)=E(y|x)$

In practice r is replaced by the constructed variables Ti, allwavesm and Tld.

Additional evidence can be provided by Hausman-type tests that compare estimates from the balanced and unbalanced samples. In the absence of non-response bias these estimates should be comparable, but non-response bias may affect the unbalanced and balanced samples differently, leading to a contrast between the estimates. It should be noted that the variable addition tests and Hausman-type tests may have low power; they rely on the sample of observed outcomes for y_{it} and will not capture non-response associated with idiosyncratic shocks that are not reflected in observed past health (Nicoletti 2002).

10.3 ESTIMATION

To try and allow for non-response we adopt a strategy based on the inverse probability weighted (IPW) estimator (Robins *et al.* 1995; Fitzgerald *et al.* 1998; Moffitt *et al.* 1999; Wooldridge 2002a, 2002b). This approach is grounded in the notion of missing at random or ignorable non-response (Rubin, 1976; Little and Rubin, 1987). Use r as an indicator of response (r=1 if observed, 0 otherwise) and y and x as the outcome and covariates of interest. Then:

1 *Missing completely at random (MCAR)* is defined by:

$$P(r=1|y, x)=P(r=1)$$

2 *Missing at random (MAR)* is defined by:

$$P(r=1|y, x)=P(r=1|x)$$

The latter implies that, after conditioning on observed covariates, the probability of non-response does not vary systematically with the outcome of interest. By Bayes rule, the MAR condition can be inverted to give:

$P(y|x,r=1)=P(y|x)$

This result provides a rationale for the Verbeek and Nijman (1992) approach to testing, which tests whether r has a place in the model for y, after conditioning on the observables x.

Fitzgerald *et al.* (1998) extend the notion of ignorable non-response by introducing the concepts of selection on observables and selection on unobservables. This requires an additional set of observables, z, that are available in the data but not included in the regression model for y. Selection on observables is defined by Fitzgerald *et al.* by the conditional independence condition:

$P(r=1|y, x, z)= P(r=1|x, z)$

Selection on unobservables occurs if this conditional independence assumption does not hold. Selection on unobservables, also termed informative, non-random or non-ignorable non-response, is familiar in the econometrics literature where the dominant approach to non-response follows the sample selection model (Heckman 1976; Hausman and Wise 1979). This approach relies on the z being 'instruments' that are good predictors of non-response and that satisfy the exclusion restriction $P(y|x,z)=P(y|x)$. This is quite different from the selection on observables approach that seeks *z's* that are endogenous to y. It is worth mentioning that linear fixed effects panel estimators are consistent, in the presence of selection on unobservables, so long as the non-ignorable non-response is due to time-invariant unobservables (see e.g., Verbeek and Nijman 1992).

The validity of the selection on observables approach hinges on whether the conditional independence assumption holds and non-response can be treated as ignorable, once z is controlled for. If the condition does hold, consistent estimates can be obtained by weighting the observed data by the inverse of the probability of response, conditional on the observed covariates (Robins *et al.* 1995). This gives more weight to individuals who have a high probability of non-response, as they are under-represented in the observed sample.

Fitzgerald et al. (1998) make it clear that this approach will be applicable when interest centres on a structural model for $P(y|x)$ and that the *z's* are deliberately excluded from the model, even though they are endogenous to the outcome of interest. They suggest lagged dependent variables as an obvious candidate for z. Rotnitzky and Robins (1994) offer a similar interpretation when they describe possible candidates for z being intermediate variables in the causal pathway from x to y. This property implies that it would not be sensible to use solely 'field variables' such as changes in interviewer as candidates for the additional observables. These kinds of variables may be good predictors of non-response but are unlikely to be associated with SAH. Horowitz and

Manski (1998) show that if the observables *(z)* are statistically independent of *y,* conditional on (x, r=1), then the weighted estimates reduce to the unweighted ones. This would explain why no difference between weighted and unweighted estimates may be reported in empirical analyses that use inappropriate variables for z.

In our application we are interested in the distribution of health problems conditional on socioeconomic status, rather than the distribution conditional on socioeconomic status and on other indicators of morbidity. We use past morbidity among our z variables. Of course, this approach will break down if an individual suffers an unobserved health shock, that occurs after their previous interview, that leads them to drop out of the survey and that is not captured by conditioning on lagged measures of morbidity. In this case non-response would remain non-ignorable even after conditioning on z. It is possible to test the validity of the selection on observables approach. The first step is to test whether the *z's* do predict non-response; this is done by testing their significance in probit models for non-response at each wave of the panel. The second is to do Hausman-type tests to compare the coefficients from the weighted and unweighted estimates. In addition the probit models for health problems can be compared in terms of the magnitudes of estimated partial effects.

Implementation of the Fitzgerald *et al.* (1998) form of the ignorability condition implies that x is observable when r=0. In the case of the kind of unit non-response we are dealing with in the BHPS, non-response means that there are missing data for the current period covariates *(x)* as well as health problems *(y)*. So we implement a stronger form of conditional independence $-P(r=1|y,x,z)=P(r=l|z)-$ as proposed by Wooldridge (2002a). To compute the IPW estimator we estimate (probit) equations for response (r_{it} =1) versus non-response (r_{it}=0) at each wave, t=2,…, T, conditional on a set of characteristics (z_{il}) that are measured for all individuals at the first wave. As described above, this relies on selection on observables and implies that non-response can be treated as ignorable non-response, conditional on Z_{i1} (Fitzgerald *et al.* 1998; Wooldridge 2002b, p. 588). Selection on observables requires that Z_{i1} contains variables that predict non-response and that are correlated with the outcome of interest but which are deliberately excluded from the model for health.

In practice z_{i1} includes the initial values of all of the regressors in the health equation. Also it includes initial values of hprob and of the other indicators of morbidity. In addition, Z_{i1} includes initial values of the respondent's activity status, occupational socioeconomic group and region. The following code is used to create variables that contain the initial values of the regressors at wave 1:

```
• sort pid wavenum
  • for each X of varlist male widowed nvrmar divsep deghdeg
  hndalev ocse hhsize nch04
  nch511 nch1218 age age2 age3 nonwhite
  selfemp unemp retired matleave famcare student ltsick
  prof mantech skillmn ptskill unskill armed
  lninc hlghq1 hprob{
  by pid: gen 'X' t1='X' [1]
  }
```

These are included in a global variable list:

```
• global z1 “malet1 widowedt1 nvrmart1 divsept1
    deghdegt1 hndalevt1 ocset1 hhsizet1 nch04t1
    nch511t1 nch1218t1 aget1 age2t1 age3t1 nonwhitet1
    selfempt1 unempt1 retiredt1 matleavet1 famcaret1
    studentt1 ltsickt1 proft1 mantecht1 skillmnt1 ptskillt1
    unskillt1 armedt1 lninct1 hlghq1t1 hprobt1 sexzero
    sfazero spozero svpzero”
```

These variables are used in a sequence of probit models for response versus non-response: so the dependent variable is insampm, which indicates whether an observation is in the estimation sample at each wave. The probits are estimated at each wave of the panel, from wave 2 to wave 8, using the full sample of individuals who were observed at wave 1. The whole loop is executed quietly as its purpose is just to create the new variable ipw: the inverse of the fitted probability of responding. For the purposes of illustration we have shown how the inverse Mills ratios (imr) could also be created and saved if this procedure was being used to do Heckman-type sample selection correction:

```
• forvalues j=2 (1) 8 {
    quietly probt insampm $z1 if (wavenum == ‘j’)
    predict p ‘j’, p
    predict lc ‘j’, xb
    gen imr ‘j’ = normden (1c ‘j’) / normprob (lc ‘j’)
    gen ipw ‘j’ =1/p ‘j’
    dprobit insampm $z1 if (wavenum == ‘j’)
    }

• gen imr=0
    • forvalues k=2(1)8 {
    replace imr=imr ‘k’ if wavenum == ‘k’
    }

• gen ipw=1
    • forvalues k=2(1) 8 {
    replace ipw=ipw ‘k’ if wavenum == ‘k’
    }

• sum ipw imr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ipw	79487	1.338138	.9675544	1	168.3076
imr	79487	.3600614	.2231152	0	2.836962

The summary statistics show that the weights (ipw) vary from 1 to 168. The inverse of the fitted probabilities from these models, $1/\hat{p}_{it}$, are then used to weight observations in the IPW-ML estimation of the pooled probit model using:

$$LogL = \sum_{i}^{n}\sum_{t}^{T}(R_{it}/\hat{p}_{it})LogL_{it}$$

Wooldridge (2002a) shows that, under the ignorability assumption:

$P(r_{it}=1|y_{it}, x_{it}, z_{i1})=P(r_{it}=1|z_{i1}), t=2,..., T$

the IPW-ML estimator is $\sqrt{n}$-consistent and asymptotically normal. Wooldridge (2002a) also shows that using the estimated $\hat{p}_{it}$ rather than the true $p_{it,}$ and ignoring the implied adjustment to the estimated standard errors leads to 'conservative inference' so that the standard errors are larger than they would be with an adjustment for the use of fitted rather than true probabilities (see also Robins *et al.* 1995). The results presented below use the unadjusted standard errors produced by Stata.

This IPW-ML estimator is implemented by adding the pweight subcommand to the probit model. The dynamic pooled probit model, with Mundlak-Wooldridge specification, is estimated for the unbalanced and balanced samples (Tables 10.1 and 10.2). These results can be compared with the unweighted estimates from the previous chapter:

• dprobit$yvar$xvarw [pweight=ipw], robust cluster (pid)

Table 10.1 Dynamic pooled probit with IPW, unbalanced panel

Probit regression, reporting marginal effects — Number of obs= 51711
Wald chi2 (25)= 8967.55
Prob > chi2= 0.0000
Log pseudolikelihood== −14393.014 — Pseudo R2= 0.3961
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>|z|	x-bar		[95% C.I.]
hprobt 1*	.444875	.0100472	54.24	0.000	.161214	.425183	.464567
male*	.0081284	.0040841	1.99	0.046	.463894	.000124	.016133
widowed*	−.0052575	.0067503	−0.77	0.442	.103249	−.018488	.007973
nvrmar*	.0093355	.0066208	1.44	0.150	.154098	−.003641	.022312
divsep*	.0125101	.0075296	1.72	0.086	.071049	−.002248	.027268
deghdeg*	−.0228032	.0069586	−3.07	0.002	.098799	−.036442	−.009165
hndalev*	−.019895	.0054129	−3.53	0.000	.206467	−.030504	−.009286
ocse*	−.0214462	.0049208	−4.22	0.000	.277063	−.031091	−.011802
hhsize	−.0018695	.0023222	−0.81	0.421	2.72003	−.006421	.002682
nch04	−.0123259	.00552	−2.23	0.026	.136402	−.023145	−.001507

nch511	−.002719	.0040232	−0.68	0.499	.249411	−.010604	.005166
nch1218	−.0011617	.0045206	−0.26	0.797	.167162	−.010022	.007698
age	.0159579	.002942	5.42	0.000	48.6175	.010192	.021724
age2	−.0279745	.0057821	−4.84	0.000	27.0309	−.039307	−.016642
age3	.0165876	.0035526	4.68	0.000	16.6738	.009625	.02355
nonwhite*	.0271869	.0119963	2.43	0.015	.039894	.003675	.050699
lninc	−.0020625	.0040842	−0.51	0.614	9.4846	−.010067	.005942
prof*	−.0572387	.0075202	−5.75	0.000	.032199	−.071978	−.042499
mantech*	−.0530807	.0044366	−10.36	0.000	.176344	−.061776	−.044385
ski llmn*	−.0648623	.0045232	−11.24	0.000	.11855	−.073728	−.055997
ptskill*	−.0474186	.0049993	−7.95	0.000	.084968	−.057217	−.03762
unskill*	−.0469891	.0073639	−5.18	0.000	.025919	−.061422	−.032556
armed*	−.0906157	.0189409	−2.00	0.045	.000461	−.127739	−.053492
hprobt1*	.1841619	.0089354	25.91	0.000	.132379	.166649	.201675
mlninc	−.0295239	.0058688	−5.02	0.000	9.48094	−.041026	−.018021
obs. P	.1732006						
pred. P	.1081427 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1 z and P>|z| correspond to the test of the underlying coefficient being 0

• dprobit $yvar $xvarw [pweight=ipw] if allwavesm==1, robust cluster (pid)

Table 10.2 Dynamic pooled probit with IPW, balanced panel

Probit regression, reporting marginal effects — Number of obs= 41879
Wald chi2 (24)= 6802.21
Prob > chi2= 0.0000
Log pseudolikelihood=−11207.173 — Pseudo R2= 0.3948
(standard errors adjusted for clustering on pid)

hprob	dF/dx	Robust Std. Err.	z	P>z|	x-bar	[95% C.I.]	
hprobt_1*	.4530882	.011408	49.40	0.000	.151484	.430729	.475447
male*	.0060932	.004394	1.39	0.165	.456556	−.002519	.014705
widowed*	−.0052489	.0073364	−0.70	0.481	.093539	−.019628	.00913
nvrmar*	.0090749	.0071951	1.29	0.197	.142469	−.005027	.023177
divsep*	.0053312	.0079006	0.69	0.493	.069307	−.010154	.020816
deghdeg*	−.0185125	.0076195	−2.29	0.022	.101329	−.033447	−.003578
hndalev*	−.0127328	.0058518	−2.12	0.034	.213319	−.024202	−.001264
ocse*	−.0156683	.0053006	−2.88	0.004	.282817	−.026057	−.005279
hhsize	−.0038159	.002525	−1.51	0.131	2.74553	−.008765	.001133
nch04	−.0141903	.0059121	−2.40	0.016	.139994	−.025778	−.002603
nch511	−.006309	.0043526	−1.45	0.147	.256781	−.01484	.002222
nch1218	.0006284	.0047349	0.13	0.894	.169187	−.008652	.009909
age	.0148146	.0032823	4.51	0.000	48.3831	.008381	.021248

age2	−.0262364	.006504	−4.04	0.000	26.5721	−.038984	−.013489
age3	.0153606	.0040338	3.81	0.000	16.1223	.007454	.023267
nonwhite*	.0232792	.0142537	1.75	0.080	.034367	−.004658	.051216
lninc	−.0061728	.0044655	−1.38	0.167	9.50933	−.014925	.002579
prof*	−.0605512	.007022	−6.05	0.000	.032974	−.074314	−.046788
mantech*	−.0496041	.0046202	−9.27	0.000	.18206	−.058659	−.040549
skillmn*	−.0577127	.0047915	−9.54	0.000	.119115	−.067104	−.048322
ptskill*	−.0462254	.0051539	−7.38	0.000	.085091	−.056327	−.036124
unskill*	−.0387779	.0078448	−4.12	0.000	.026875	−.054153	−.023402
hprobt 1*	.1821808	.0099687	23.42	0.000	.123304	.162643	.201719
mlninc	−.0283436	.0064659	−24.38	0.000	9.50677	−.041017	−.015671
obs. P	.1615294						
pred. P	.0988768 (at x-bar)						

(*) dF/dx is for discrete change of dummy variable from 0 to 1 z and P>|z| correspond to the test of the underlying coefficient being 0

A comparison of these results with their unweighted equivalents shows some differences in the estimated partial effects, but these tend to be very small in magnitude.

The IPW-ML estimator can be adapted to allow the elements of *z* to be updated and change across time, for example adding *z* variables measured at *t–1* to predict response at *t.* This should improve the power of the probit models to predict non-response and hence make the ignorability assumption more plausible. In this case the probit model for non-response at wave *t* is estimated relative to the sample that is observed at wave *t-1.* This relies on non-response being an absorbing state and is therefore confined to 'monotone attrition', where respondents never re-enter the panel. Also, because estimation at each wave is based on the selected sample observed at the previous wave, the construction of inverse probability weights has to be adapted. The predicted probability weights are constructed cumulatively using $\hat{p}_{it} = \hat{\pi}_{i2} \times \hat{\pi}_{i2} \ldots \times \hat{\pi}_{it}$, where the $\hat{\pi}_{it}$ denote the fitted selection probabilities from each wave. In this version of the estimator the ignorability condition has to be extended to include future values of *y* and *x* (see Wooldridge 2002b, p. 589). Once again Wooldridge shows that omitting a correction to the asymptotic variance estimator leads to conservative inference.

The IPW approach is attractive as it is easy to apply in the context of non-linear models, such as the probit model, and only requires a reweighting of the data. In contrast to the published longitudinal weights that are supplied with the BHPS, our IPW weights are model-specific and specifically designed for the outcome of interest and the associated problem of health-related non-response, although the validity of the approach depends on the credibility of the ignorability assumption.

Jones, Koolman and Rice (2006) show that there is clear evidence of health-related non-response in both the BHPS and ECHP. In general, individuals in poor initial health are more likely to drop out, although for younger groups non-response is associated with good health. Furthermore, variable addition tests provide evidence of non-response bias in the models of SAH. Nevertheless a comparison of estimates based on the balanced samples, the unbalanced samples and models corrected for non-response using inverse probability weights shows that, in many cases, substantive differences in the magnitudes

of the average partial effects of lagged health, income and education are small. Similar findings have been reported concerning the limited influence of non-response bias in models of income dynamics and various labour market outcomes and on measures of social exclusion such as poverty rates and income inequality indices.

11
Models for health-care use

11.1 INTRODUCTION

Many empirical analyses of the use of health-care services use as dependent variable a count variable (non-negative integer valued count *y=0, 1, ...*) such as the number of visits to a physician (sometimes detailed by type of physician), number of hospital stays or number of drug prescriptions. In the recent literature there are various examples of empirical modelling of count measures of health care such as Grootendorst (1995; number of different medicines used), Pohlmeier and Ulrich (1995; visits to GP, visits to specialist), Hakkinen *et al.* (1996; visits to doctor), Gerdtham (1997; physician visits, weeks in hospital), Deb and Trivedi (1997; physician office visits, nonphysician office visits, physician hospital outpatient visits, non-physician hospital outpatient visits, emergency room visits and hospital stays), Santos Silva and Windmeijer (2001; specialist visits, GP visits) and Deb and Trivedi (2002; number of outpatient visits). The data on health-care utilization typically contain a large proportion of zero observations, as well as a long right tail of individuals who make heavy use of health care. The basic count data regression model is the Poisson. This model has been shown to be too restrictive for modelling health-care utilization, and more general specifications have been preferred.

This chapter illustrates the use of count data models. The models are applied to Portuguese data taken from waves 2 to 4 of the European Community Household Panel (ECHP), covering the years 1995 to 1998. The dependent variable is the number of visits to a specialist in the previous 12 months.

Empirical studies of health utilization usually consider as regressors variables that measure: need/morbidity (more commonly, self assessed status, with four or five categories, but also indicators of chronic conditions and limited activity, days of sickness/restricted activity and, ideally, albeit not usually available in survey data, objective health measures); age (accounting for imperfect health status measurement but also for individual preferences); sex (accounting for gender-specific health-care requirements and also for tastes); ability to pay (income, wealth) and other socio-demographic factors such as marital status, education level attained, labour market status and job characteristics. Some studies have also considered the price of health care and characteristics of insurance coverage and, less commonly, owing to lack of data, time costs and accessibility. In this chapter we consider an admittedly small list of covariates, as the main goal is to illustrate the practical issues involved in the various methodologies. When interpreting the results, we should therefore bear in mind that the estimated effects may be capturing effects of omitted variables. The explanatory variables considered here are: age in years, a dummy variable for sex (male), the logarithm of household income (lh in come) and a dummy variable that equals one if lagged self-assessed health status is bad or very bad (lsahbad). The ECHP income variable is total net household income. Here,

this variable is deflated by national consumer price indices (CPI), making it comparable across the panel, and by purchasing power parities (PPP), which would have allowed for comparability across countries. The income variable was further deflated by the OECD modified equivalence scale in order to account for household size and composition. In order to make the syntax in the remainder of this chapter more general, it is useful to create a global list of regressors:

• global xvar "age male lhincome lsahbad"

11.2 THE POISSON MODEL

The basic count data model is the Poisson model. The dependent variable, y_i, is assumed to follow a Poisson distribution, with mean λ_i, defined as a function of the covariates X_i. Thus, the model is defined by:

$$P(y_i) = e^{-\lambda i}\lambda_i^{yi}/y_i!$$

where the conditional mean λ_i is usually defined as:

$\lambda_i=E(y_i|x_i)=exp(x_i\beta)$

The poisson regression model is estimated for the ECHP data on the number of visits to the specialist (labelled y) and predictions are saved for both the fitted values $exp(x_i\beta)$ and the linear index $x_i\beta$:

• poisson y $xvar
  • predict fitted, n
  • predict yf, xb

Table 11.1 contains the results of maximum likelihood estimation of the Poisson regression model. The output contains the estimated coefficients, standard errors and resulting z-ratios for each explanatory variable. The coefficients relate to the linear index $x_i\beta$, while the expected number of visits is a non-linear function of the *x's*. Thus, the β*'s* are not measured in the original units of the count data and inferences about the effect of a given variable on the number of doctor visits require the re-transformation of coefficient estimates. The coefficients can nevertheless be used to analyse the qualitative impacts of the variables considered. In line with the findings of previous analyses of health-care utilization, the results show positive effects of age, income and poor health, and a negative effect of being male.

Table 11.1 Poisson regression for number of specialist visits

Poisson regression				Number of obs=		32164
				LR chi2 (4)=		9782.96
				Prob>chi2=		0.0000
Log likelihood=−66421.47				Pseudo R2=		0.0686
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0023812	.0003307	7.20	0.000	.001733	.0030295
male	−.3564767	.0107427	−33.18	0.000	−.3775321	−.3354214
lhincome	.3606307	.0078924	45.69	0.000	.3451618	.3760996
lsahbad	.9008154	.0120738	74.61	0.000	.8771512	.9244795
_cons	−3.212611	.072741	−44.17	0.000	−3.35518	−3.070041

The marginal effect of a continuous explanatory variable x_k, is given by the formula:

$\partial E(y_i/x_i)/\partial x_{ik}=\beta_k exp(x_i\beta)$

The average effect of a binary variable is given by:

$E(y_i|x_{ik}=1)-E(y_i|x_{ik}=0)=exp(x_i\beta|x_{ik}=1)-exp(x_i\beta|x_{ik}=0)$

As these marginal and average effects depend on the value of the remaining explanatory variables, it is common to evaluate them at the sample means of the other regressors. Alternatively, estimates can be calculated for every observation. For example, we can compute the effect of male on the expected number of specialist visits, and then take the sample average:

- scalar bmale=_b [male]
  - gen ae_male=0
  - replace ae_male=exp (yf+bmale)−exp (yf) if male==0
  - replace ae_male=exp (yf)−exp (yf−bmale) if male==1

The estimated partial effects can be summarized and plotted using a histogram, although for brevity the results are not presented here:

- summ ae_male
  - hist ae_male

The performance of the model can be assessed by the tabulation of actual against fitted values of *y*. These fitted values are rounded to the nearest integer:

- replace fitted=round (fitted)
  - tab y fitted
  - drop fitted

The RESET test of correct specification can be performed in the usual way:

• gen yf2=yf^2
  • quietly poisson y $xvar yf2
  • test yf2
  • drop yf2

In this case, the output of the RESET test is:

```
(1) [y] yf2=0
           chi2 (1)=102.02
         Prob>chi2=0.0000
```

indicating strong evidence against the null hypothesis of correct specification of the Poisson model.

The Poisson model implies equality of the conditional mean and conditional variance. This is called the equidispersion property and it has been shown to be too restrictive in many empirical applications. In case of over- or underdispersion, the maximum likelihood estimator will still give consistent estimates of β. However, the resulting estimates of the standard errors are biased.

As an alternative approach, an appeal to the Poisson pseudo-maximum likelihood estimator (PMLE) can be used. The estimator for β is defined by the first-order conditions of the MLE but the distribution need not be Poisson. In other words, the Poisson mean assumption is maintained but the restriction of equidispersion is relaxed. This is done by using an alternative estimator for the covariance matrix (different functional forms can be assumed for the conditional variance of y_i; see, for example, Cameron and Trivedi (1998)). The option robust specifies that the covariance matrix should be estimated using the Huber-White sandwich estimator:

• poisson y $xvar, robust

Table 11.2 shows the results of the Poisson pseudo-maximum likelihood estimation. The coefficient estimates result from maximum likelihood estimation, so they are the same as above, while the standard errors result from the Huber-White sandwich estimator.

Table 11.2 Poisson regression for number of specialist visits with robust standard errors

Poisson regression				Number of obs=		32164
				Wald chi2 (4)=		1594.92
				Prob>chi2=		0.0000
Log pseudolikelihood=−66421.47				Pseudo R2=		0.0686
y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0023812	.0009238	2.58	0.010	.0005706	.0041919
male	−.3564767	.0318036	−11.21	0.000	−.4188107	−.2941428
lhincome	.3606307	.0228243	15.80	0.000	.3158958	.4053656
lsahbad	.9008154	.0367396	24.52	0.000	.8288072	.9728236
_cons	−3.212611	.2122474	−15.14	0.000	−3.628608	−2.796613

The literature on modelling of health-care utilization has shown that the Poisson model is usually too restrictive. This has motivated the use of different parametric distributions that can account for the features of the data that are inconsistent with the Poisson. Cameron and Trivedi (1998) list the most common departures from the standard Poisson model. Some of these deal with problems that often arise when modelling count measures of health-care utilization such as: failure of the equidispersion property (usually accounted for by considering a mixture model for the unobserved heterogeneity); 'excess zeros' problem (higher observed frequency of zeros than is consistent with the Poisson); and multimodality (if observations are drawn from different populations, the observed distribution can be multimodal). The remainder of this chapter covers generalizations of the Poisson model that have been used to overcome its limitations for modelling health-care utilization.

11.3 THE NEGATIVE BINOMIAL MODEL

Cameron and Trivedi (1998) note that one of the reasons for the failure of the Poisson regression is unobserved heterogeneity. Neglected unobserved heterogeneity leads to overdispersion and excess of zeros. The heterogeneity can be modelled as a mixture, by considering $exp(x_i\beta+\mu_i)=[exp_ix_i\beta)\eta_i$, with $E(\eta_i)=1$ and η_i a random term whose distribution should be defined. While in the Poisson model it is considered that $(y_i|x_i)$ follows a Poisson distribution, in the mixture model the Poisson distribution is assumed for $(y_i|x_i,\eta_i)$. Defining the distribution of η_i, leads to the marginal distribution of $(y_i|x_i)$. The Negative Binomial model (NB) can be derived as a Poisson mixture where the η_i is gamma distributed (for the derivation, see, for example, Cameron and Trivedi, 1998). The associated probability of observing the count y_i is then:

$$P(y_i)=\{\Gamma(y_{i+}\psi_i)/\Gamma(\psi_i)\Gamma(y_i+1)\}\ (\psi_i/(\lambda_i+\psi_i))^{\psi i}(\lambda_i/(\psi_i))^{yi}$$

where $\Gamma(.)$ is the gamma function. Considering $\psi=(1/\alpha)\lambda^k$, for $a>0$, gives:

$$E(y)=\lambda \text{ and } Var(y)=\lambda+\alpha\lambda^{2-k}$$

The NB model nests the Poisson model, which is given when a=0. Most empirical applications of the NB model consider k=1 or k=0 (NB1 and NB2). In the NB1 the variance is proportional to the mean, $(1+\alpha)\lambda$, while in the NB2 the variance is a quadratic function of the mean, $\lambda+\alpha\lambda^2$. By default, Stata estimates the NB2 model. We save the vector of estimated coefficients, the number of *x's* (including constant) and the estimation results for later use:

- nbreg y $xvar
- matrix bnb=e (b)
- scalar k=colsof (bnb)−1
- estimates store lcnb2

The estimation results of the NB2 model are shown in Table 11.3. The conditional mean function is defined in the same way as in the Poisson model, so the coefficients should be interpreted in the same way. Additionally, the estimate for the over-dispersion parameter αequals 3.46 and is highly significant. This means that the equidispersion property (imposed by the Poisson model) is rejected. Despite this the estimated coefficients show only small differences, compared to the Poisson model, while the estimated standard errors and t-ratios are substantially different.

Table 11.3 Negative Binomial model for number of specialist visits

Negative binomial regression				Number of obs=		32164
				LR chi2 (4)=		1830.26
				Prob>chi2=		0.0000
Log likelihood=−42753.001				Pseudo R2=		0.0210
y	Coef.	Std. Err.	z	P>|z|	[95%Conf.	Interval]
age	.0064446	.0007321	8.80	0.000	.0050097	.0078795
male	−.4560705	.0238621	−19.11	0.000	−.5028394	−.4093016
lhincome	.30484	.0158709	19.21	0.000	.2737336	.3359463
lsahbad	.8853403	.0286717	30.88	0.000	.8291449	.9415358
_cons	−2.893422	.1444795	−20.03	0.000	−3.176596	−2.610247
/lnalpha	1.241131	.0146867			1.212345	1.269916
alpha	3.459523	.0508091			3.361358	3.560554

Likelihood-ratio test of alpha=0: chibar2 (01)=4.7e+04 Prob>=chibar2= 0.000

Following the estimation of the model, we can calculate partial effects and fitted values in the same way as for the Poisson:

- predict fitted, n
- predict yf, xb
- scalar bmale=_b [male]
- gen ae_male=0

- replace ae_male=exp (yf+bmale)−exp (yf) if male==0
- replace ae_male=exp (yf)−exp (yf−bmale) if male==1
- summ ae_male
- hist ae_male
- drop ae_male yf

- replace fitted=round (fitted)
- tab y fitted
- drop fitted

The alternative NB1 specification can be obtained by using the option dispersion (constant). A generalization of the NB2 model is obtained allowing α to vary with the regressors. In particular, *log(α)* is parameterized as a linear combination of the regressors.

- gnbreg y $xvar, lna($xvar)
- predict fitted
- replace fitted=round (fitted)
- tab y fitted
- drop fitted

Table 11.4 shows significant coefficients for all covariates in the overdispersion equation. All the variables have estimated coefficients with opposite signs on the conditional mean function and on the overdispersion function.

Table 11.4 Generalized Negative Binomial model for number of specialist visits

Generalized negative binomial regression				Number of obs=		32164
				LR chi2 (4)=		1571.10
				Prob>chi2=		0.0000
Log likelihood=−42307.817				Pseudo R2=		0.0182
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
age	.0040104	.0007655	5.24	0.000	.0025101	.0055107
male	−.4007285	.0254426	−15.75	0.000	−.450595	−.3508619
lhincome	.3771582	.0171184	22.03	0.000	.3436068	.4107096
lsahbad	.8455891	.027503	30.75	0.000	.7916842	.899494
_cons	−3.406211	.1585513	−21.48	0.000	−3.716966	−3.095456
lnalpha						
age	−.0071049	.0009375	−7.58	0.000	−.0089424	−.0052673
male	.5215486	.030414	17.15	0.000	.4619383	.5811589
lhincome	−.4503186	.0209182	−21.53	0.000	−.4913176	−.4093196
lsahbad	−.389146	.0340555	−11.43	0.000	−.4558936	−.3223985
_cons	5.436645	.1920486	28.31	0.000	5.060237	5.813053

According to Gurmu (1997) 'although the NB model is superior to the Poisson in that it allows for overdispersion, it is inadequate in various practical situations'. Gurmu notes that there is evidence of poor fit in counts models with excess zeros and long-tailed distributions. The assumption that the zeros and positive observations are generated by the same process has been shown to be too restrictive in the case of health-care utilization. Pohlmeier and Ulrich (1995), who model the number of visits to a doctor, argue that the decision of first contact and the frequency of visits are determined by two different processes. While the first can be considered to depend only on the individual, the frequency of visits also reflects supply characteristics and the influence of providers. The different nature of the zeros and the positive observations has been taken into account by two alternative approaches: the zero inflated models and the hurdle models. These specifications are presented below.

11.4 ZERO INFLATED MODELS

The zero inflated model gives more weight to the probability that the count variable equals zero. It incorporates an underlying mechanism that splits individuals between non-users, with probability $q(x_{1i}\beta_1)$, and potential users, with probability $1-q(x_{1i}\beta_1)$. The probability function for the zero-inflated Poisson model, $P^{ZIP}(y|x)$, is a mixture of the standard Poisson model, $P^{p}(y|x)$ and a degenerate distribution concentrated at zero:

$$P^{ZIP}(y|x)=1(y=0)q+(1-q)P^{P}(y|x)$$

A more general specification is obtained when the NB model, instead of the Poisson, is used for the number of visits of the potential users, ZINB. Zero inflated Poisson and NB models can be estimated by maximum likelihood. The simplest version is the ZIP with constant zero-inflation probability q. If the estimation command includes the option vuong, then the Vuong statistic is displayed with the estimation results, which allows the comparison of the non-nested ZIP and Poisson models.

- zip y $xvar, inflate (_cons) vuong
- predict fitted
- predict yf, xb
- replace fitted=round (fitted)
- tab y fitted
- drop fitted

As can be seen in Table 11.5, the Vuong test of ZIP against the Poisson model clearly favours the zero inflated specification. This shows evidence of a split between users and non-users of specialist visits. The estimated results for the Poisson model allowing for zero inflation are substantially different from the ones obtained previously with the basic Poisson model.

Table 11.5 Zero Inflated Poisson model for number of specialist visits I

Zero-inflated poisson regression				Number of obs=		32164
				Nonzero obs=		11266
				Zero obs=		20898
Inflation model = logit				LR chi2 (4)=		1900.46
Log likelihood=−51057.84				Prob>chi2=		0.0000
y	Coef.	Std. Err.		P>| z |	[95% Conf.	Interval]
y						
age	−.0007622	.000364	−2.09	0.036	−.0014757	−.0000487
male	−.070633	.0118146	−5.98	0.000	−.0937892	−.0474768
lhincome	.0947423	.0083167	11.39	0.000	.0784419	.1110427
lsahbad	.5153494	.0125093	41.20	0.000	.4908315	.5398672
_cons	.199003	.0776417	2.56	0.010	.0468281	.3511779
inflate						
cons	.5093962	.0124118	41.04	0.000	.4850695	.5337229

Vuong test of zip vs. standard Poisson: z=37.55 Pr>z=0.0000

Computation of partial effects needs to be different from what was shown above for the Poisson and NB models, in order to account for the different mean function in the ZIP model:

- scalar bmale=_b [male]
  - scalar qi=_b [inflate:_cons]
  - scalar qi=exp (qi)/(1+exp (qi))
  - scalar list qi
  - gen ae_male=0
  - replace ae_male=(1−qi)*(exp (yf+bmale)−exp (yf)) if male==0
  - replace ae_male=(1−qi)*(exp (yf)−exp (yf−bmale)) if male==1
  - summ ae_male
  - hist ae_male
  - drop ae_male yf

The model can be extended to allow for the zero-inflated probability *(q)* to depend on the explanatory variables. However, researchers often report problems in getting the estimates to converge when the full set of regressors is included in the splitting mechanism (see, e.g., Grootendorst (1995) and Gerdtham (1997)).

- zip y $xvar, inflate ($xvar _cons)
- predict fitted
- replace fitted=round (fitted)
- tab y fitted
- drop fitted

The results in Table 11.6 show evidence that the split between potential users and non-users of specialist visits is influenced by all the covariates included in the model. Older individuals, as well as those with higher incomes and poorer health, tend to have a lower probability of being non-users, while the opposite is observed for males.

Table 11.6 Zero Inflated Poisson model for number of specialist visits II

Zero-inflated poisson regression				Number of obs=		32164
				Nonzero obs=		11266
				Zero obs=		20898
Inflation model=logit				LR chi2 (4)=		1617.98
Log likelihood=−50028.94				Prob>chi2=		0.0000
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
age	−.0016061	.0003606	−4.45	0.000	−.0023127	−.0008994
male	−.0246478	.0115466	−2.13	0.033	−.0472788	−.0020168
lhincome	.0628027	.0080689	7.78	0.000	.0469879	.0786174
lsahbad	.4836837	.0124048	38.99	0.000	.4593707	.5079967
_cons	.5299433	.0747959	7.09	0.000	.383346	.6765406
inflate						
age	−.0087677	.000786	−11.16	0.000	−.0103081	−.0072272
male	.6159362	.0255089	24.15	0.000	.5659397	.6659327
lhincome	−.5068374	.0195151	−25.97	0.000	−.5450863	−.4685885
lsahbad	−.7045381	.0317606	−22.18	0.000	−.7667878	−.6422884
_cons	5.232266	.177915	29.41	0.000	4.883559	5.580973

Estimation of the ZINB can be done using with similar commands (Table 11.7):

- zinb y $xvar, inflate (_cons)
- predict fitted
- replace fitted=round (fitted)
- tab y fitted
- drop fitted

Table 11.7 Zero Inflated NB model for number of specialist visits I

Zero-inflated negative binomial regression				Number of obs=		32164
				Nonzero obs=		11266
				Zero obs=		20898
Inflation model=logit				LR chi2 (4)=		1800.70
Log likelihood=−42753				Prob>chi2=		0.0000
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
age	.0064445	.0007321	8.80	0.000	.0050096	.0078794
male	−.4560676	.0238619	−19.11	0.000	−.5028361	−.4092991
lhincome	.3048394	.0158708	19.21	0.000	.2737332	.3359455
lsahbad	.8853373	.0286714	30.88	0.000	.8291425	.9415322
_cons	−2.893434	.1444785	−20.03	0.000	−3.176606	−2.610261
inflate						
_cons	−16.88837	645.133	−0.03	0.979	−1281.326	1247.549
/lnalpha	1.241124	.0146868	84.51	0.000	1.212339	1.26991
alpha	3.459501	.050809			3.361336	3.560532

Similarly to what was done in the ZIP, the zero-inflation probability can be parameterized as a function of the explanatory variables:

- zinb y $xvar, inflate ($xvar _cons) vuong
  - predict fitted
  - replace fitted=round (fitted)
  - tab y fitted
  - drop fitted

The results for the ZINB with parameterized zero-inflation probability are shown in Table 11.8. The estimated α is highly significant, which shows evidence against the nested ZIP. On the other hand, the Vuong test favours the ZINB against the NB without zero-inflation. The estimated coefficients in the NB model for the potential users differ significantly from the ones obtained with the NB regression without zero-inflation (Table 11.3). For example, the simpler specification indicated a negative and significant effect of being male on the expected number of specialists visits. In the ZINB, the negative coefficient of male on the number of visits of potential users is not significant, while there is evidence that males have a substantially larger probability of being non-users.

Table 11.8 Zero Inflated NB model for number of specialist visits II

Zero-inflated negative binomial regression				Number of obs=		32164
				Nonzero obs=		11266
				Zero obs=		20898
Inflation model=logit				LR chi2 (4)=		507.28
Log likelihood=−42218.81				Prob>chi2=		0.0000
y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
age	−.0043025	.0008231	−5.23	0.000	−.0059159	−.0026892
male	−.027775	.0284726	−0.98	0.329	−.0835804	.0280304
lhincome	.2112907	.018254	11.58	0.000	.1755135	.247068
lsahbad	.6247848	.0296696	21.06	0.000	.5666334	.6829361
_cons	−1.347752	.174639	−7.72	0.000	−1.690038	−1.005465
inflate						
age	−.0426388	.0031538	−13.52	0.000	−.0488201	−.0364576
male	1.816027	.1017853	17.84	0.000	1.616532	2.015523
lhincome	−.598092	.0437903	−13.66	0.000	−.6839195	−.5122646
lsahbad	−2.279397	.2978691	−7.65	0.000	−2.863209	−1.695584
_cons	5.282505	.4004838	13.19	0.000	4.497572	6.067439
/lnalpha	.8565125	.0281352	30.44	0.000	.8013685	.9116566
alpha	2.354934	.0662566			2.228589	2.488441

Vuong test of zinb vs. standard negative binomial: z=15.09 Pr>z=0.0000

11.5 HURDLE MODELS

The hurdle model implies that the count measure of health-care utilization is a result of two different decision processes. The first part specifies the decision to seek care, and the second part models the positive values of the variable for those individuals who receive some care. This can be interpreted as a principal-agent type model, where the physician (the agent) determines utilization on behalf of the patient (the principal) once initial contact is made. Thus, it is assumed that the decision to seek care is taken by the individual, while the level of care depends also on supply factors.

It has been shown in the literature on health-care utilization that the two-part hurdle model is often a better starting point than the NB class (for example, Pohlmeier and Ulrich 1995; Grootendorst 1995; Gerdtham 1997). Another motivation for the hurdle model is the high proportion of zeros that remains even after allowing for overdispersion.

The hurdle model for count data was proposed by Mullahy (1986). The participation decision and the positive counts are determined by two different processes $P_1(.)$ and $P_2(.)$. The log-likelihood for the hurdle model is given by:

$$LogL=\Sigma_{y=0}\, log[1-P_1(y>0|x)]+\Sigma_{y>0}\{log[P_1(y>0|x)]$$
$$+log[P_2(y|x\,y>0)]\}$$
$$=\{\Sigma_{y=0}\, log[1-P_1(y>0|x)]+\Sigma_{y>0}\, log[P_1(y>0|x)]\}$$
$$+\{\Sigma_{y>0}\, log[P_2(y|x,y>0)]\}$$
$$=LogL_1+LogL_2$$

The two parts of the hurdle model can be estimated separately. For the participation decision, a binary model has to be defined. The second decision gives the amount of use of health care, given participation, and is modelled by a truncated at zero count data model. There have been several applications of the hurdle model in the context of health-care utilization. The distribution P_1 is usually logit, probit, Poisson or NB, while the most common choices for P_2 are the Poisson and the NB.

In Mullahy (1986), the underlying distribution for both stages is the Poisson. Pohlmeier and Ulrich argue that it is necessary to account for remaining unobserved heterogeneity, since 'supply-side effects are rarely well captured in household data at the micro level'. Thus, these authors use the NB1 distribution for both stages of the model, instead of the Poisson. They note that this specification allows for explicit testing of distributional assumptions (for example, against Poisson) and the equality of the two parts of the decision-making process (thus, assessing the importance of considering that the number of physician visits is determined by two different processes). The NB (with k=0, 1) is the most used distribution in empirical applications of the hurdle model to the utilization of health care.

Gurmu (1997) notes a possible practical problem related to the hurdle model. When the sample size is small or the proportion on zeros is very high, it might be difficult to estimate the second part of the model. Gurmu suggests that, in this case, the researchers should focus on modelling the first stage, using binary models. Another problem related to the estimation of the second part of the hurdle model should be noted - the decision depends on supply characteristics that are generally unobserved.

The following commands request the estimation of a logit model for the probability that the number of visits is positive and save the vector of estimated coefficients, blogit, the value of the maximized log-likelihood, logl_logit, and the number of observations, N, for later use (Table 11.9).

• gen biny=y>0
  • logit biny $xvar
  • drop biny
  • matrix blogit=e (b)
  • scalar logl_logit=e (11)
  • scalar N=e (N)

Table 11.9 Logit model for the probability of having at least one visit to a specialist

Logit estimates				Number of obs=	32164
				LR chi2 (4)=	2334.78
				Prob>chi2=	0.0000
Log likelihood=−19662.402				Pseudo R2=	0.0560
biny	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
age	.0081008	.0007538	10.75	0.000	.0066234 .0095782
male	−.6061245	.0246338	−24.61	0.000	−.6544058 −.5578432
lhincome	.51122	.0188779	27.08	0.000	.47422 .54822
lsahbad	.7879607	.0308416	25.55	0.000	.7275122 .8484091
_cons	−5.348587	.1721851	−31.06	0.000	−5.686063 −5.01111

A truncated at zero NB2 is used for the second part of the model. This model is estimated over the observations with positive y (Table 11.10):

- ztnb y $xvar if y>0
  - matrix btrunc=e (b)
  - scalar logl_trunc=e (11)

Table 11.10 Truncated at zero NB2 for the number of specialist visits

0-Truncated Negative Binomial Estimates				Number of obs=	11266
				Model chi2(4)=	419.94
				Prob>chi2=	0.0000
Log Likelihood=−22613.5884191				Pseudo R2=	0.0092
y	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
y					
age	−.0008151	.0008804	−0.93	0.355	−.0025407 .0009106
male	−.0487244	.029037	−1.68	0.093	−.1056359 .0081872
lhincome	.0627813	.0190673	3.29	0.001	.0254102 .1001525
lsahbad	.6155092	.0314509	19.57	0.000	.5538665 .6771519
_cons	−.1872319	.1784201	−1.05	0.294	−.536929 .1624651
lnalpha					
_cons	.7778782	.0566325	13.74	0.000	.6668806 .8888758

alpha 2.176848 [lnalpha]_cons=ln(alpha)

(LR test against Poisson, chi2 (1)=87615.76 P=0.0000)

The hurdle model is composed of the two parts in Tables 11.9 and 11.10. Table 11.9 shows a positive effect of age, income and poor health and a negative effect of male on

the probability of visiting a specialist. Table 11.10 shows that, conditional on having at least one visit, the expected number of visits increases significantly with poor health and income. The negative effect of male is only significant at 10%, while the effect of age is non-significant. The estimated overdispersion parameter is again highly significant.

The log-likelihood of the hurdle model is obtained from the sum of those of the logit and the truncated NB2. Additionally, the Akaike and Schwarz information criteria (AIC and BIC, respectively) are computed and displayed for comparison with models estimated in the remainder of this chapter.

- scalar logl_hurdle=logl_logit+logl_trunc
- scalar aic=−2*log1_hurdle+2*(2*k+1)
- scalar bic=−2*log1_hurdle+log (N)*(2*k+1)
- display “loglhurdle=” logl_hurdle
  “aic hurdle=” aic “bic hurdle=” bic

loglhurdle=−42275.99 aic_hurdle=84573.981 bic_hurdle= 84666.145

11.6 FINITE MIXTURE/LATENT CLASS MODELS

Deb and Trivedi (1997) propose the use of finite mixture models as an alternative to the hurdle models in the empirical modelling of health-care utilization. In a more recent paper Deb and Trivedi (2002) point out that ‘a more tenable distinction for typical cross-sectional data may be between an “infrequent user” and a “frequent user” of medical care, the difference being determined by health status, attitudes to health risk, and choice of lifestyle’. They argue that this is a better framework than the hurdle model, and that it distinguishes more starkly between users and non-users of care.

Deb and Trivedi (1997) point out a number of advantages of the finite mixture approach. It provides a natural representation since each latent class can be seen as a ‘type’ of individual, while still accommodating heterogeneity within each class. It can also be seen as a discrete approximation of an underlying continuous mixing distribution that does not need to be specified. Furthermore, the number of points of support needed for the finite mixture model is low, usually two or three.

In the finite mixture (latent class) formulation of unobserved heterogeneity, the latent classes are assumed to be based on the person’s latent long-term health status, which may not be well captured by proxy variables such as self-perceived health status and chronic health conditions (Cameron and Trivedi 1998). The two-point finite mixture model suggests the dichotomy between the ‘healthy’ and the ‘ill’ groups, whose demands for health care are characterized by, respectively, low mean and low variance and high mean and high variance. Jimenez-Martin *et al.* (2002) agree with the advantages of the finite mixture model described above but also note a disadvantage. Namely, while the hurdle model is a natural extension of an economic model (the principal-agent model), the finite mixture model is driven by statistical reasoning.

In a latent class (LC) model the population is assumed to be divided into C distinct populations in proportions $\pi_1,\ldots\pi_C$, where $\sum_{j=1}^{C} \pi_j = 1$, $0 \le \pi_j \le 1$, $j=1,\ldots, C$. The C-point finite mixture model is given by:

$$f(y_i|.) = \sum_{j=1}^{C} \pi_j . f_j(y_i|.),$$

where the mixing probabilities, π_j, are estimated along with all the other parameters of the model. Also, $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$. The component distributions in a C-point finite mixture negative binomial model are defined as:

$$f_j(y_i|.) = \{\Gamma(y_i+\psi_{j,i})|\Gamma(\psi_{j,i})\Gamma(y_i+1)\}(\psi_{j,i}|(\lambda_{j,i}+\psi_{j,i}))^{\psi j,i}(\lambda_{j,i}|(\lambda_{j,i}+\psi_{j,i}))^{yi}$$

where $j=1,\ldots C$ are the latent classes $\lambda_{j,i}=exp(x_i\beta_j)$ *and* $\psi_{j,i} = (1/\alpha_j)\lambda_{j.i}^{k}$. In the most general specification, all the elements of the vectors β_j are allowed to vary across the latent classes. However, more parsimonious specifications can arise from restrictions on the components of (α_j, β_j). For example, the slope parameters can be restricted to be equal across all latent classes. In this case, the differences between classes are given only by differences in the intercept.

The number of classes in a finite mixture model is commonly chosen according to information criteria, such as the AIC and BIC. First, the Negative Binomial model without any mixture is estimated, followed by the LC model with C=2. The number of points of support C is chosen such that there is no further improvement in the information criteria when C is increased. In most applications C equals 2 or 3. In this chapter, we illustrate the estimation of the LCNB with C=2, but the procedures can easily be extended to accommodate a larger number of latent classes.

Since there is no built-in command in Stata for the estimation of the latent class NB, it is necessary to define a program to do this estimation. The following code defines a latent class model with two latent classes, using an NB2 for each latent class:

```
• program define lcnb2
    version 8.0
    args lnf b1 a1 b2 a2 bpi
    tempvar f_1 f_2 pi
    gen double ‘f_1’=0
    gen double ‘f_2’=0
    gen double ‘pi’=0
    quietly replace ‘pi’=exp (‘bpi’)/(1+exp (‘bpi’))
    quietly replace ‘f_1’=lngamma (y+1/‘a1’))
    1/‘a1’) −lngamma(
              −lngamma(y+1)−1/ ‘a1’*log(1+‘a1’*exp(‘b1’))
              −y*log(1+exp(−‘b1’)/‘a1’)
    quietly replace ‘f_2’=lngamma (y+1/‘a’)
    (1/‘a2’)−lngamma
```

```
            −lngamma(y+1)−1/‘a2’*log(1+‘a2’*exp(‘b2’))
            −y*log(1+exp(−‘b2’)/‘a2’
    quietly replace ‘lnf’=
            log(‘pi’*exp(‘f_1’)+(1−‘pi’)*exp(‘f_2’))
end
```

The probability of belonging to class 1 equals π (represented in the program lcnb2 by the temporary variable pi). Accordingly, the probability of belonging to class 2 equals $(1-\pi)$. In order to ensure that the class membership probabilities fall between 0 and 1, π is parameterized using the logistic function. Thus, we estimate the log-odds-ratio $log(1-\pi)$). The program lcnb2 can easily be adapted to the finite mixture of the NB1 distribution.

Owing to the possibility of convergence to local maxima in mixture models, the estimation should be repeated using different sets of starting values for the parameters being estimated. These starting values can be obtained as combinations of the estimates of the one component version of the model, saved in the vector bnb. Here, the starting values for both classes are equal to the estimates of nbreg, except for the constant terms, which are defined as the constant term of nbreg multiplied by (1−dif_init) and (1+dif_init). In order to start the estimation with π_0= 0.5, the starting value for the log-odds ratio is 0.

- scalar dif_init=.20
- mat initc1= (bnb[1, 1..k−1], (1−dif_init)*bnb[1, k], exp(bnb[1, k+1]))
- mat initc2= (bnb[1, 1..k−1], (1+dif_init)*bnb[1, k], exp(bnb[1, k+1]))
- mat initlcnb=(initc1, initc2, 0)

It is possible to modify the vector bnb in a number of different ways such as: (i) modifying more parameters than just the constant terms; (ii) choosing different values for dif_init; (iii) adding and subtracting a constant instead of multiplying by a modifying factor; etc. Alternatively, estimates of restricted versions of the latent class model (for example, with constant slopes) can be used as starting values in the estimation of more flexible versions. The model is estimated by maximum likelihood using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton algorithm. The following commands are given to estimate the model specified by the program lcnb2, departing from the parameter values contained in initlcnb:

- ml model If lcnb2 (xb1: $xvar) (alfa1:) (xb2: $xvar) (alfa2:) (pi:), technique(bfgs)
- ml init initlcnb, skip copy
- ml maximize, nooutput
- ml display, diparm(pi, invlogit p)

When nooutput is specified as an option in ml maximize, Stata suppresses the display of the final results and shows just the iteration log. The output can then be specified using the options of the command ml display. The option diparm (pi, invlogit p) determines that the displayed table of results is to include not only the estimate for $log(\pi/(1-\pi))$ (constant term in equation pi) but also π. This produces the output presented in Table 11.11: xb1 and alfa1 contain the parameters of the NB2 for class 1, xb2 and alf a2 correspond to

class 2, and /pi gives the estimated probability of belonging to class 1. The estimated class proportions are 0.319 and 0.681. All the variables have coefficients of the same sign in both classes but they are all larger in absolute value, and more significant, for class 2.

Table 11.11 LCNB2 model for the number of specialist visits (with two latent classes)

					Number of obs=	32164
					Wald chi2(4)=	279.55
Log likelihood=−42411.029					Prob>chi2=	0.0000
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb1						
age	.003028	.0011167	2.71	0.007	.0008393	.0052166
male	−.1926734	.036409	−5.29	0.000	−.2640338	−.121313
lhincome	.0810925	.0271553	2.99	0.003	.0278692	.1343158
lsahbad	.6691618	.0437118	15.31	0.000	.5834882	.7548354
_cons	−.053742	.26351	−0.20	0.838	−.5702121	.462728
alfal						
_cons	1.615614	.1311003	12.32	0.000	1.358662	1.872566
xb2						
age	.0113378	.0020644	5.49	0.000	.0072916	.015384
male	−1.078371	.0839858	−12.84	0.000	−1.24298	−.9137617
lhincome	1.150551	.0861978	13.35	0.000	.9816069	1.319496
lsahbad	1.425018	.0815161	17.48	0.000	1.265249	1.584786
_cons	−11.38746	.8201293	−13.88	0.000	−12.99488	−9.780036
alfa2						
_cons	2.692809	.1252221	21.50	0.000	2.447378	2.93824
pi						
_cons	−.7584398	.123921	−6.12	0.000	−1.001321	−.5155591
/pi	.3189851	.0269198	11.85	0.000	.2686819	.3738912

We can compute fitted values for each latent class and analyse their summary statistics (Table 11.12):

- predict xb1, eq(xb1)
- predict xb2, eq(xb2)
- gen y_c1=exp (xb1)
- gen y_c2=exp (xb2)
- sum y_c1 y_c2
- drop xb1 xb2 y_c1 y_c2

The mean fitted value and the minimum value are substantially larger for class 1. The maximum fitted value for class 2 is larger than for class 1, but y_c1 is larger than y_c2

for all but 1% of the observations. Therefore we refer to class 1 and class 2 as high users and low users, respectively.

Table 11.12 Summary statistics of fitted values by latent class (LCNB2)

Variable	Obs	Mean	Std. Dev.	Min	Max
y_c1	32164	2.499076	.9450703	1.220708	5.64865
y_c2	32164	.6029967	.9684484	.0007493	37.57004

The equality of the coefficients of all covariates across latent classes can be tested using a Wald test:

• test [xb1=xb2]

The output shows clear rejection of the null hypothesis:

```
(1) [xb1]age−[xb2]age=0
(2) [xb1]male−[xb2]male=0
(3) [xb1]lhincome−[xb2]lhincome=0
(4) [xb1]lsahbad−[xb2] lsahbad=0
           chi2(4)=248.98
         Prob>chi2= 0.0000
```

It has already been noted that the estimated coefficients are larger in absolute value for class 2 (low users). Tests of equality of coefficients of individual covariates across classes can be performed in order to test whether the coefficients are significantly larger for low users. For example, for income:

• test [xb1]lhincome=[xb2]lhincome

There is clear evidence that the income coefficient is larger for low users:

```
(1) [xb1]lhincome−[xb2]lhincome=0
           chi2(1)=133.50
         Prob>chi2=0.0000
```

Similar results are obtained for the other three regressors.

The hurdle and the LCNB models are usually compared using information criteria (AIC and BIC). As noted above, these criteria are also used to choose the number of latent classes. We display these after estimation together with those stored above for the NB2:

• estimates store lcnb2
• estimates stats nb2 lcnb2

Table 11.13 shows that the AIC and the BIC improve considerably when two latent classes are considered, instead of the one component NB2 model. These AIC and BIC for the LCNB2 are, however, substantially larger than those shown above for the hurdle model, which means that the latter specification is preferred according to these criteria.

Table 11.13 AIC and BIC of NB2 and LCNB2 (with two latent classes) for the number of specialist visits

Model	nobs	ll (null)	ll (model)	df	AIC	BIC
nb2	32164	−43668.13	−42753	6	85518	85568.27
lcnb2	32164	.	−42411.03	13	84848.06	84956.98

In practice, we should now move on to a model with three latent classes and set C equal to the number beyond which the information criteria do not improve.

Recent empirical studies of health-care utilization have provided comparisons between the performance of the hurdle model and the latent class model. In the empirical applications in Deb and Trivedi (1997, 2002), it is found that a two-point mixture of NB is sufficient to explain health-care counts very well and that it outperforms the NB hurdle. Deb and Holmes (2000) also present evidence that the finite mixture model outperforms the hurdle model. Jimenez-Martin *et al.* (2002), however, show that, in some cases, the hurdle model can provide better results than the finite mixture model. They compare the hurdle and the finite mixture specifications for visits to specialists and GPs in 12 EU countries. It is found that the finite mixture model performs better for visits to GPs while the hurdle model is preferred for visits to specialists.

11.7 LATENT CLASS HURDLE MODEL

Bago d'Uva (2006) proposes a model that combines the hurdle and the finite mixture models in a single specification. Drawing on the latent class model, the unobserved individual heterogeneity is represented by a finite number of classes. Then, for each class, the hypothesis that the decision concerning the number of visits is taken in two steps is not discarded. Individual health-care use in a given period is therefore assumed to be determined by a two-stage decision process, conditional on the latent class.

The combination of the hurdle and the latent class framework in a cross-sectional context poses identification problems, arising from the non-identifiability of the finite mixture of the binary model with a single outcome. Bago d'Uva (2006) specifies a latent class hurdle for panel data (LCH-Pan), which considers the panel structure in the formulation of the mixture. In the LCH-Pan, the latent class framework represents individual unobserved time-invariant heterogeneity. In other words, the distribution of the individual effects is approximated by a discrete distribution. Furthermore, the model accommodates heterogeneity in the slopes, as these can be allowed to vary across latent classes.

Recent empirical studies have used the latent class framework to model binary indicators of health-care utilization in a panel data context (or with multiple binary

responses in a cross section). Atella *et al.* (2004) model the probability of visiting three types of physician jointly. The individuals are assumed to be drawn from a population with latent classes. Within each latent class, the decision to visit each physician type follows a probit distribution. An example of a binary mixture model with panel data is the discrete random effects probit. Deb (2001) uses a latent class model where only the intercept varies across classes. The discrete random effects probit is a discrete approximation of the distribution of the unobserved family effects in the random effects probit. Bago d'Uva (2005) uses the latent class approach to account for individual unobserved heterogeneity in panel data models for access to and utilization of primary care. Conditional on the latent class, it is assumed that the probability of visiting a GP in a given year is determined by the logit model. In the model for the number of GP visits, as the information on the dependent variable is grouped, an aggregated NB is used for each latent class.

There are a number of applications of latent class models in other fields (e.g. Wang *et al.* 1998; Wedel *et al.* 1993; Nagin and Land 1993; Uebersax 1999). Greene (2001) notes that most applications have not used panel data. However, according to Greene, the latent class model is 'only weakly identified at very best by a cross section'. Additionally, he notes that the richness of the panel in terms of cross-group variation improves the potential for estimating the model. The recent implementation of latent class models for panel data in LIMDEP 8.0 (Greene 2002) suggests that this approach may become more popular in the near future (for counts, LIMDEP 8.0 contains built-in commands for the estimation of latent class Poisson and NB models). In the context of smoking behaviour, Clark and Etilé (2003) use the latent class framework to approximate the continuous distribution of the individual effects in a dynamic random effects bivariate probit model. Clark *et al.* (2005) develop a latent class ordered probit model for reported well-being, in which individual time-invariant heterogeneity is allowed both in the intercept and in the income effect.

This chapter uses a panel of individuals across time. Individuals i are observed T_i times. Let y_{it} represent the number of visits in year t. Denote the observations of the dependent variable over the panel as $y_i=[y_{i1},\ldots, y_{iTi}]$. Consider that individual i belongs to a latent class j, $j=1,\ldots$, C, and that individuals are heterogeneous across classes. Conditional on the covariates considered, there is homogeneity within a given class j. Given the class that individual i belongs to, the dependent variable in a given year t, y_{it}, has density $f_j(y_{it}|x_{it}, \theta_j)$ and the θ_j are vectors of parameters that are specific to each class. The joint density of the dependent variable over the observed periods is a product of T_i independent densities $f_j(y_{it}|x_{it}, \theta_j)$, given class j. The probability of belonging to class j is π_{ij}, where $0<\pi_{ij}<1$ and $\sum_{j=1}^{C}\pi_{ij}=1$. Unconditionally on the latent class the individual belongs to, the joint density of $y_i=[y_{i1},\ldots, y_{iTi}]$ is given by:

$$g(y_i \mid x_i; \pi_{i1}, \ldots, \pi_{iC}; \theta_1, \ldots, \theta_C) = \sum_{j=1}^{C} \pi_{ij} \prod_{t=1}^{T_i} f(y_{it} \mid x_{it}; \theta_j)$$

where $\mathbf{x}_i$ is is a vector of covariates, including a constant, and θ_j are vectors of parameters.

The estimation of latent class models for panel data in Stata also requires the definition of specific programs. As these models assume that each individual belongs to the same latent class throughout the panel, it is useful to convert the data from the usual long form (where each row represents one period *t* for individual *i,* identified here by variable pidc, and each column represents one variable *z*) to wide form (where each row represents one individual, and variable *z* across periods 1,…, *T* is represented by columns *z1*,… *zT):*

• reshape wide y $xvar, i (pidc) j (wave)

The output displayed after this command describes clearly the transformations that reshape creates in the dataset:

(note: j = 1 2 3 4)

Data	long	->	wide
Number of obs.	32164	->	8041
Number of variables	8	->	22
j variable (4 values)	wave	->	(dropped)
xij variables:			
	y	->	y1 y2 … y4
	age	->	age1 age2 … age4
	male	->	male1 male2 … male4
	lhincome	->	lhincome1 lhincome2 … lhincome4
	lsahbad	->	lsahbad1 lsahbad2 … lsahbad4

New lists of variables are created to be used in the estimation of latent class models for panel data:

• global xvar1 “age1 male1 lhincome1 lsahbad1”
• global xvar2 “age2 male2 lhincome2 lsahbad2”
• global xvar3 “age3 male3 lhincome3 lsahbad3”
• global xvar4 “age4 male4 lhincome4 lsahbad”

Conditionally on the class that the individual belongs to, the number of visits in period *t,* y_{it}, is assumed to be determined by a hurdle model. The underlying distribution for the two stages of the hurdle model is the negative binomial. Formally, for each component j=1, … *C,* it is assumed that the probability of zero visits and the probability of observing y_{it} visits, given that y_{it} is positive, are given by the following expressions:

$$f_j(0 \mid x_{it};\, \beta_{j1}) = P[y_{it} = 0 \mid x_{it},\, \boldsymbol{\beta}_{j1}] = (\alpha_j \lambda_{j1,it}^{k} + 1)^{-\frac{1}{\alpha_j}}$$

$$f_j(y_{it} \mid y_{it} > 0,\, x_{it};\, \boldsymbol{\beta}_{j2}) = \frac{\Gamma\left(y_{it} + \frac{\lambda_{j2,it}^{k}}{\alpha_j}\right)(\alpha_j \lambda_{j2,it}^{k} + 1)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}} \left(1 + \frac{\lambda_{j2,it}^{k-1}}{\alpha_j}\right)^{-y_{it}}}{\Gamma\left(\frac{\lambda_{j2,it}^{k}}{\alpha_j}\right)\Gamma(y_{it}+1)\left[1 - (\alpha_j \lambda_{j2,it}^{k} + 1)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}}\right]}$$

where $\lambda_{j1,it} = \exp(x'_{it}\beta_{j1}),\ \lambda_{j2,it} = \exp(x'_{it}\beta_{j2}),$ α_j are overdispersion parameters and k is as defined above. So, in this case, $\theta_j=(\beta_j,\alpha_j)$. The same set of regressors is considered in both parts of the model.

Having $[\beta_{j1},\beta_j 2] \neq [\beta_{l1},\beta_{l2}]$ for $j \neq l$ reflects the differences between the latent classes. It can be assumed that all the slopes are the same, varying only the constant terms, $\beta_{j1,0}$ and $\beta_{j2,0}$, and the overdispersion parameters α_j. This represents a case where there is unobserved individual heterogeneity but not in the responses to the covariates (as in the model used in Deb (2001)). The most flexible version allows α_j and all elements of β_{j1} and β_{j2} to vary across classes.

On the other hand, similarly to the hurdle model, the fact that β_{j1} can be different from β_{j2} reflects the possibility that the zeros and the positives are determined by two different decision processes. In other words, the determinants of care are allowed to affect differently the two stages of the decision process regarding the number of visits to the doctor—the probability of seeking care and the number of visits, given that this is positive. The finite mixture hurdle model accommodates a mixture of sub-populations for which health-care use is determined by an NB model (the two decision processes are indistinguishable), and sub-populations for which utilization is determined by a hurdle model. This is obtained by setting $\beta_{j1}=\beta_{j2}$, for some classes. If those restrictions are imposed in all classes, then we have a finite mixture NB for panel data, FMNB-Pan. This model differs from the LCNB (Deb and Trivedi 1997, 2002) presented above, in that it accounts for the panel structure of the data. Comparison of the non-nested models LCNB and FMNB-Pan shows the extent to which it is relevant to account for the panel data structure in the latent class framework.

In most empirical applications of latent class models to health-care utilization, class membership probabilities are taken as fixed parameters $\pi_{ij}=\pi_j$, j=1, …, C, to be estimated along with $\boldsymbol{\theta}_1$, …, $\boldsymbol{\theta}_c$ (Deb and Trivedi 1997, 2002; Deb and Holmes 2000; Deb 2001; Jimenez-Martin *et al.* 2002, Atella *et al.* 2004). This corresponds to the assumption that the individual heterogeneity is uncorrelated with the regressors, also inherent in random effects or random parameters specifications.

Let us start by illustrating the estimation of the FMNB-Pan with constant class membership probabilities. Similarly to what was done above for the LCNB2, program lcnb2_pan defines the log-likelihood of a model with two latent classes and an NB2 for each class. The program is specific to a balanced panel with four waves. Temporary variables f_j (j=1,2) represent the logarithm of $\prod_{t=1}^{T_i} f_j(y_{it} \mid x_{it};\theta_j)$, that is, the joint density of the dependent variable over the observed periods, where the density for each period is

NB2. The specification of the model in this way requires that the dataset is converted to wide form, which we have done above:

- capture program drop lcnb2_pan
- program define lcnb2_pan
  version 8.0

```
args lnf b1_w1 a1_w1 b2_w1 a2_w1 bpi
            b1_w2 a1_w2 b1_w3 a1_w3 b1_w4 a1_w4
            b2_w2 a2_w2 b2_w3 a2_w3 b2_w4 a2_w4
   tempvar f_1 f_2 pi
   gen double ‘f_1’=0
   gen double ‘f_2’=0
   gen double ‘pi’=0
   quietly replace ‘pi’=exp (‘bpi’) / (1+exp (‘bpi’))
   quietly replace ‘f_1’=lngamma (y1+1/ ‘a1_w1’)
            −lngamma(1/‘a1_w1’)−lngamma(y1+1)
            −1/‘al_wl’*log(1+‘a1_w1’*exp (‘b1_w1’))
            −y1*log(1+exp(−‘b1_w1’)/‘a1_w1’)
               +lngamma(y2+1/‘a1_w2’)
            −lngamma(1/‘a1_w2’)−lngamma(y2+1)
            −1/‘a1_w2’*log(1+‘a1_w2’*exp (‘b1_w2’))
            −y2*log(1+exp(−‘b1_w2’)/‘a1_w2’)
               +lngamma(y3+1/‘a1_w3’)
            −lngamma(1/‘a1_w3’)–lngamma(y3+1)
            −1/‘a1_w3’ *log(1+ ‘a1_w3’ *exp(‘b1_w3’))
            −y3*log(1+exp(−‘b1_w3’)/‘a1_w3’)
               +lngamma(y4+1/‘a1_w4’)
            −lngamma(1/‘a1_w4’)−lngamma(y4+1)
            −1/‘a1_w4x’ *log(1+ ‘a1_w4’ *exp(‘b1_w4’))
            −y4*log(1+exp(−‘b1_w4’)/‘a1_w4’)
   quietly replace ‘f_2’=lngamma (y1+1/ ‘a2_w1’)
            −lngamma(1/‘a2 w1’)−lngamma(y1+1)
            −1/‘a2_1’ ‘*log (1+ ‘a2_w’ *exp (‘b2_w’))
            −y1*log (1+exp (−‘b2_w1’) / ‘a2_w1’)
               +lngamma(y2+1/‘a2_w2’)
            −lngamma(1/‘a2_w2’)−lngamma(y2+1)
            −1/‘a2_w2’ *log(1+ ‘a2_w2’ *exp(‘b2_w2’))
            −y2*log(1+exp(−‘b2_w2’)/‘a2_w2’)
               +lngamma(y3+1/‘a2_w3’)
            −lngamma(1/‘a2_w3’)—lngamma(y3+1)
            −1/‘a2_w3’*log(1+‘a2_w3’*exp (‘b2_w3’))
            −y3*log(1+exp(−‘b2_w3’)/‘a2_w3’)
               +lngamma(y4+1/‘a2_w4’)
            −lngamma(1/‘a2_w4’)−lngamma(y4+1)
            −1/‘a2_w4’ *log(1+ ‘a2_w4’ *exp(‘b2_w4’))
```

```
        −y4*1og(1+exp(−‘b2_w4’)/‘a2_w4’)
quietly replace ‘lnf’ =
        log( ‘pi’ *exp( ‘f_1’)+(1−‘pi’) *exp ( ‘f_2’))
end
```

The program lcnb2_pan does not account for the assumption of the LCNB-Pan that the parameters contained in θ_j, j=1,2, are constant throughout the panel, which has to be done through the specification of constraints to be imposed in the estimation of the model:

```
• const drop _all
• global i=1
• f oreach wave in 2 3 4{
    foreach var in $xvar {
        const $i [xb1] ‘var‘ 1=[xb1_w‘wave’] ‘var’ ‘wave’
        global i=$i+1
        const $i [xb2] ‘var‘1=[xb2_w‘wave’] ‘var’ ‘wave’
        global i=$i+1
    }
    const $i [xb1] _cons= [xb1_w‘wave’] _cons
    global i=$i+1
    const $i [xb2]_cons=[xb2_w‘wave’]_cons
    global i=$i+1
    const $i [alfa1]_cons=[alfa1_w‘wave’]_cons;
    global i=$i+1
    const $i [alfa2]_cons=[alfa2_w‘wave’]_cons;
    global i=$i+1
}
• global i=$i−1
```

Starting values initc1, initc2 and initpi are defined in the same way as for the LCNB2. For the LCNB2-Pan, we also need to initialize the parameters corresponding to waves 2 to 4, constrained to be the same as the ones for wave 1. We use initc1, initc2 as vectors of initial values for waves 1 to 4:

```
• scalar dif_init=.20
• mat initc1=
(bnb[1,1..k−1], (1−dif_init)*bnb[1,k],exp(bnb[1,k+1]))
• mat initc2 =
(bnb[1,1..k−1], (1+dif_init)*bnb[1,k],exp (bnb [1,k+1]))
• scalar initpi=0
• mat initlcpan=(initc1, initc2, initpi,
        initc1,initc1,initc1, initc2, initc2, initc2)
```

With the syntax below, the LCNB2-Pan is estimated and the parameters of interest are displayed. The option const (1-$i) imposes the specified constraints during estimation. As above, the output is not displayed upon convergence (ml maximize, nooutput). Instead,

only the estimation results for the first five equations (xb1, alfa1, xb2, alfa2 and pi) are displayed. The omitted results correspond to the parameters for waves 2 to 4, restricted to be the same as for wave 1. As for the LCNB2, the option diparm (pi, invlogit p) requests that the estimated π be displayed:

- ml model If lcnb2_pan (xb1; $xvar1) (alfa1:)
  (xb2: $xvar1) (alfa2:) (pi;)
(xb1_w2: $xvar2) (alfa1_w2) (xb1_w3: $xvar3) (alfa1_w3:)
(xb1_w4: $xvar4) (alfa1_w4:)
(xb2_w2: $xvar2) (alfa2_w2:) (xb2_w3: $xvar3) (alfa2_w3:)
(xb2_w4: $xvar4) (alfa2_w4:), technique(bfgs) const (1−
$i)
- ml init initlcpan, skip copy
- ml maximize, nooutput
- ml display, neq(5) diparm(pi,invlogit p)

Table 11.14 shows the estimation results for the LCNB2-Pan. It is interesting to compare these with the estimates of the LCNB2 in Table 11.12. The parameters are more precisely estimated in the panel data model (except for the overdispersion parameter of class 2), especially in the case of class proportions and regressor coefficients in class 1. All coefficients have the same signs as in the LCNB2. The magnitudes of the effects of male, income and poor health in class 2 are substantially larger in the pooled LCNB2. The panel model provides a better fit to the data, with the same number of parameters.

Table 11.14 LCNB2-Pan for the number of specialist visits (with two latent classes)

					Number of obs=	8041
					Wald chi2 (0)=	.
Log likelihood=−41061.181					Prob>chi2=	.
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb1						
age1	.0028428	.0009119	3.12	0.002	.0010554	.0046301
male1	−.1875673	.0305507	−6.14	0.000	−.2474456	−.127689
lhincome1	.1510011	.0180487	8.37	0.000	.1156263	.1863759
lsahbad1	.6146377	.0299932	20.49	0.000	.555852	.6734233
_cons	−.616775	.1701776	−3.62	0.000	−.9503169	−.2832331
alfa1						
_cons	1.246992	.0374341	33.31	0.000	1.173622	1.320361
xb2						
age1	.017472	.0019129	9.13	0.000	.0137227	.0212213
male1	−.6840054	.0616785	−11.09	0.000	−.8048931	−.5631178
lhincome1	.514931	.0545562	9.44	0.000	.4080029	.6218591
lsahbad1	.5055809	.0734572	6.88	0.000	.3616076	.6495543
_cons	−6.300164	.5233497	−12.04	0.000	−7.32591	−5.274417

alfa2 _cons	4.744733	.2623357	18.09	0.000	4.230564	5.258901
pi _cons	−.5839086	.0530392	−11.01	0.000	−.6878636	−.4799536
/pi	.3580337	.0121908	29.37	0.000	.3345085	.3822631

Stata also displays the list of constraints imposed, not shown here.

In order to compute fitted values, we predict $x_{it}\beta_j$ for each latent class and each wave, then reshape back to long form and compute fitted values and the respective summary statistics:

```
• predict xb1_1, eq(xb1)
  • predict xb2_1, eq(xb2)
  • foreach wave in 2 3 4 {
  predict xb1_‘wave’, eq (xb1_w ‘wave’)
  predict xb2_‘wave’, eq (xb2_w ‘wave’)
  }
  reshape long y $xvar xb1_xb2_, i (pidc) j (wave)
  gen y_c1=exp (xb1_)
  gen y_c2=exp (xb2_)
  drop xb1_xb2_
  sum y_c1 y_c2
  drop y_c1 y_c2
```

The mean fitted value, maximum and minimum values are substantially larger for class 1, to which we can refer as class of high users (Table 11.15). The disparity between the mean fitted values in the LCNB2-Pan is larger than what was shown in Table 11.12 for the LCNB2.

Table 11.15 Summary statistics of fitted values by latent class (LCNB2-Pan)

Variable	Obs	Mean	Std. Dev.	Min	Max
y_c1	32164	2.521107	.8735386	.9542556	6.536865
y_c2	32164	.3579982	.2416224	.0150148	3.213232

In order to assess to what extent the two classes respond differently to the covariates considered, we can perform tests of equality of slopes. When considered jointly, the responses of the two classes of users are significantly different:

```
• test [xb1=xb2]
  (1) [xb1]age1− [xb2]age1= 0
  (2) [xb1]male1− [xb2]male1= 0
  (3) [xb1]lhincome1− [xb2]lhincome1= 0
  (4) [xb1]lsahbad1− [xb2]lsahbad1= 0
            chi2(4) =      146.53
```

```
Prob > chi2 =          0.0000
```

In particular, the estimated effect of income is significantly higher for low users:

```
• test [xb1]lhincome1= [xb2]lhincome1
   (1) [xb1]lhincome1− [xb2]lhincome1= 0
               chi2(1) =     44.73
   Prob > chi2 =          0.0000
```

The same conclusion applies to the coefficients of male and age. No significant difference is found between the estimated coefficients of lsahbad for high and low users:

```
• test [xb1]lsahbad1= [xb2]lsahbad1
   (1) [xb1]lsahbad1− [xb2]lsahbad1= 0
         chi2(1) =          1.91
   Prob > chi2=          0.1668
```

The LCNB2-Pan can be compared with the one-component NB2 and the LCNB2 by means of the information criteria AIC and BIC:

```
• estimates store lcnb2_pan
   • estimates stats nb2 lcnb2 lcnb2_pan
```

Table 11.16 shows that the LCNB2-Pan performs better than the NB2 according to information criteria, which provides evidence of unobserved individual heterogeneity. The panel version of the latent class model outperforms the LCNB2.

Table 11.16 AIC and BIC of NB2, LCNB2 and LCNB2-Pan (with two latent classes) for the number of specialist visits

Model	nobs	11 (null)	11(model)	df	AIC	BIC
nb2	32164	−43668.13	−42753	6	85518	85568.27
lcnb2	32164	.	−42411.03	13	84848.06	84956.98
lcnb2_pan	8041	.	−41061.18	13	82148.36	82239.26

We turn now to the LCH-Pan, allowing, within each latent class, for the possibility that the zeros and the positives are determined by two different decision processes. Again, we present a program for a model with two latent classes and four waves of data, which can be easily extended to a specification with more classes and a longer panel. The program lchurdle_pan extends lcnb2_pan by considering in the construction of temporary variables f_1 and f_2, the density of the hurdle model for each period. Similarly to the one-component hurdle model in Table 11.10, we consider a logit for the binary part and a truncated at zero NB2 for the second part. The list of arguments of the new program contains equations for the binary part of the hurdle model (b1_pr_w1 to b1_pr_w4, and b2_pr_w1 to b2_pr_w4) and for the truncated part (b1_tr_w1 to b1_tr_w4, b2_tr_w1 to

b2_tr_w4, a1_tr_w1 to a1_tr_w4 and a2_tr_w1 to a2_tr_w4). As in the lcnb2_pan, we impose the constraints that the parameters are the same throughout the panel. This program assumes that the dataset is in wide form so we start by returning to this form:

```
• reshape wide y $xvar, i (pidc) j (wave)
• capture program drop lchurdle_pan
• program define lchurdle_pan
    version 8.0
    args lnf b1_pr_w1 b1_tr_w1 a1_tr_w1
            b2_pr_w1 b2_tr_w1 a2_tr_w1
            bpi
        b1_pr_w2 b1_pr_w3 b1_pr_w4
        b2_pr_w2 b2_pr_w3 b2_pr_w4
        b1_tr_w2 a1_tr_w2 b1_tr_w3 a1_tr_w3
        b1_tr_w4 a1_tr_w4
        b2_tr_w2 a2_tr_w2 b2_tr_w3 a2_tr_w3
        b2_tr_w4 a2_tr_w4
    tempvar f_1 f_2 pi
    gen double ‘f_1’=0
    gen double ‘f_2’=0
    gen double ‘pi’=0
    quietly replace ‘pi ’=exp (‘bpi’) / (1+exp (‘bpi’))
    quietly replace ‘f_l’=(lngamma (y1+1/ ‘a1_tr_w1’)
    − lngamma(1/‘a1_tr_w1) − lngamma(y1+1)
    − log ((1+‘a1_tr_w1 ’*exp (‘b1_tr_w1’)) ^ (1 /‘a1_tr_w1’) −1)
    − y1*log(1+exp(−‘bl_tr_w1’)/‘a1_tr_w1’)) * (y1>0)
    − log (exp (‘b1_pr_w1’) +1)+‘b1_pr_w1’ * (y1>0)
            + (lngamma(y2+1/‘a1_tr_w2’)
    − lngamma(1/‘a1_tr_w2’) − lngamma(y2+1)
    − log( (1+‘a1_tr_w2’ *exp(‘b1_tr_w2’)) ^ (1/ ‘a1_tr_w2’) −1)
    − y2*log (1+exp(−‘b1_tr_w2’/‘a1_tr_w2’)) * (y2>0)
    − log (exp (‘b1_pr_w2’) +1)+‘b1_pr_w2’ * (y2>0)
            + (lngamma(y3+1/‘a1_tr_w3’)
    − lngamma(1/‘a1_tr_w3’) − lngamma(y3+1)
    − log((1+‘a1_tr_w3’ *exp(‘b1_tr_w3’)) ^(1/‘a1_tr_w3’) −1)
    − y3*log(1+exp(−‘b1_tr_w3’)/‘a1_tr_w3’)) * (y3>0)
    − log(exp(‘b1_pr_w3’) +1)+‘b1_pr_w3’ * (y3>0)
            + (lngamma (y4+1/‘a1_tr_w4’)
    − lngamma(1/‘a1_tr_w4’) − lngamma (y4+1)
    − log((1+ ‘a1_tr_w4−’ *exp(‘b1_tr_w4’)) ^ (1/‘a1_tr_w4’) −1)
    − y4*log(1+exp(−‘b1_tr_w4’)/‘a1_tr_w4’)) * (y4>0)
    − log (exp (‘b1_pr_w4’) +1)+‘b1_pr_w4’ * (y4>0)
    quietly replace ‘f_2’ = (lngamma (y1+1/‘a2_tr_w1’)
    − lngamma(1/‘b2_tr_w1’) − lngamma (y1+1)
    − log( (1+ ‘a2_tr_w1’ *exp( ‘b2_tr_w1’)) ^ (1/‘a2_tr_w1’) −1)
    − y1*log(1+exp(−‘b2_tr_w1’)/‘Za2_tr_w’)) * (y1>0)
```

```
    − log (exp (‘b2_pr_w1’) +1)+‘b2_pr_w1’ * (y1>0)
            + (lngamma(y2+l/‘a2_tr_w2’)
    − lngamma(1/‘a2_tr_w2’) − lngamma (y2+1)
    − log((1+‘a2_tr_w2’ *exp (‘b2_tr_w2’)) ^(1/‘a2_tr_w2’) −1)
    − y2*1og(1+exp(−‘b2_tr_w2’) /‘a2_tr_w2’)) * (y2>0)
    − log (exp (‘b2_pr_w2’) +1)+‘b2_pr_w2’ * (y2>0)
            + (lngamma(y3+1/‘a2_tr_w3’)
    − lngamma(1/‘a2_tr_w3’) − lngamma (y3+1)
    − log((1+‘a2_tr_w3’ *exp (‘b2_tr_w3’)) ^ (1/‘a2_tr_w3’) −1)
    − y3*log (1+exp (−‘b2_tr_w3’) / ‘a2_tr_w3’)) * (y3>0)
    − log (exp (‘b2_pr_w3’) +1)+‘b2_pr_w3’ * (y3>0)
            + (lngamma (y4+1/‘a2_tr_w4’)
    − lngamma(1/‘a2_tr_w4’) − lngamma (y4+1)
    − log((1+‘a2_tr_w4’ *exp(‘b2_tr_w4’)) ^ (1/‘a2_tr_w4’) −1)
    − y4*log (1+exp (−‘b2_tr_w4’) / ‘a2_tr_w4’)) * (y4>0)
    − log (exp (‘b2_pr_w4’) +1)+ ‘b2 pr w4’* (y4>0)
    quietly replace ‘lnf’ =
        log(‘pi’*exp(‘f_1’) + (1−‘pi’) *exp (‘f_2’))
end
• const drop _all
• global i=1
• foreach wave in 2 3 4 {
    foreach part in prob trunc {
    foreach var in $xvar {
        const $i [xb1_‘part’] ‘var’ 1
        = [xb1_‘part’_w‘wave’] ‘var’ ‘wave’
    global i=$i+1
        const $i [xb2_‘part’] ‘var’1
        = [xb2_‘part’_w‘wave’] ‘var’ ‘wave’
        global i=$i+1
        }
        const $i [xb1_‘part’]_cons=[xb1_‘part’_w‘wave’]_cons
        global i=$i+1
        const $i [xb2_‘part’]_cons=[xb2_‘part’_w‘wave’]_cons
        global i=$i+1
        }
        const $i [alfa1]_cons=[alfa1_‘wave’] _cons
        global i=$i+1
        const $i [alfa2]_cons=[alfa2_w‘wave’]_cons
        global i=$i+1
}
• const list
• global i=$i−1
```

Before estimating the model, we define initial values for the parameters, using the estimates of the hurdle model (Table 11.10). The vector initlchurdle is constructed in a

similar way as initlcpan above, except that now we need to initialize the parameters of the binary part (initc1_prob, initc2_prob) and of the truncated part (initc1_trunc, initc2_trunc). The initial value for the *α's* is defined as exp (btrunc [1, k+1]) since the command ztnb used for the second part of the hurdle model estimates *In (a)* instead of *a:*

- scalar dif_init=.2
- mat initc1_prob =
  (blogit[1,1. .k−1], (1−dif_init) *blogit [1,k])
- mat initc2_prob =
  (blogit[1,1..k−1], (1+ dif_init)*blogit[1, k])
- mat initc1_trunc=
  (btrunc[1,1..k−1],
  (1−dif_init)*btrnc[1,k], exp(btrunc[1,k+1]))
- mat initc2_trunc=
  (btrunc[1, 1..k−1],
  (1+dif_init)*btrunc[1,k],exp(btrunc[1,k+1]))
- mat initpi=0
- mat initlchurdle=(initc1_prob, initc1_trunc,
  initc2_prob, initc2_trunc,
  initpi,
  initc1_prob, initc1_prob, initc1_prob,
  initc2_prob, initc2_prob, initc2_prob,
  initc1_trunc, initc1_trunc, initc1_trunc,
  initc2_trunc, initc2_trunc, initc2_trunc)

As noted above, starting values can be specified in a number of different ways and estimation should be repeated with different sets of starting values in order to avoid local maxima.

The log-likelihood defined by program lchurdle_pan is maximized, starting from the vector initlchurdle. The estimation results are saved in vector blchurdle. Again, the full set of estimation results is suppressed (nooutput) and only the relevant parameters are shown (neq (6) diparm (pi, invlogit p)):

- ml model If lchurdle_pan
  (xb1_prob: $xvar1) (xb1_trunc: $xvar1) (alfa1:)
  (xb2_prob: $xvar1) (xb2_trunc: $xvar1) (alfa2:)
  (pi:)
  (xb1_prob_w2:$xvar2) (xb1_prob_w3:$xvar3)
  (xb1_prob_w4:$xvar4)
  (xb2_prob_w2:$xvar2) (xb2_prob_w3:$xvar3)
  (xb2_prob_w4: $xvar4)
  (xb1_trunc_w2:$xvar2) (alfa1_w2:)
  (xb1_trunc_w3:$xvar3) (alfa1_w3:)
  (xb1_trunc_w4:$xvar4) (alfa1_w4:)
  (xb2_trunc_w2:$xvar2) (alfa2_w2:)
  (xb2_trunc_w3:$xvar3) (alfa2_w3:)

(xb2_trunc_w4:$xvar4) (alfa2_w4:), technique(bfgs)
const(1−$i)
• ml init initlchurdle, skip copy
• ml maximize, nooutput
• ml display, neq(6) diparm(pi,invlogit p)
• matrix blchurdle=e (b)

The estimation results of the LCH-Pan are presented in Table 11.17, with class proportions estimated as 0.347 and 0.653. Consistently across classes and in both parts, positive effects are estimated for age, income and poor health, while there are negative effects for males. The coefficients in the binary part are more significant than those in the truncated part, for both classes.

Table 11.17 LCH-Pan for the number of specialist visits (with two latent classes), with constant class membership

Number of obs= 8041
Wald chi2 (0)= .
Log likelihood=−40674.74 Prob >chi2= .

	Coef.	Std. Err.	z	P>|z|	[95%Conf.	Interval]
xb1_prob						
age1	.015004	.0018359	8.17	0.000	.0114057	.0186024
male1	−.8800321	.0596411	−14.76	0.000	−.9969265	−.7631378
lhincome1	.4628983	.035815	12.92	0.000	.3927021	.5330944
lsahbadvbad1	.8976306	.0677831	13.24	0.000	.7647782	1.030483
_cons	−3.802004	.3272806	−11.62	0.000	−4.443462	−3.160545
xb1_trunc						
age1	.0017154	.0010958	1.57	0.117	−.0004323	.0038632
male1	−.0598689	.0359851	−1.66	0.096	−.1303984	.0106606
lhincome1	.0889994	.0225198	3.95	0.000	.0448613	.1331374
lsahbadvbad1	.5809995	.0373789	15.54	0.000	.5077383	.6542608
_cons	−.1328759	.2079001	−0.64	0.523	−.5403526	.2746008
alfa1						
_cons	1.641206	.0983976	16.68	0.000	1.44835	1.834062
xb2_prob						
age1	.0165937	.0015292	10.85	0.000	.0135965	.0195909
male1	−.6830616	.0541978	−12.60	0.000	−.7892874	−.5768358
lhincome1	.6353853	.043825	14.50	0.000	.5494898	.7212807
lsahbadvbad1	.4748538	.0588522	8.07	0.000	.3595057	.5902019
_cons	−7.661848	.4069399	−18.83	0.000	−8.459436	−6.864261

xb2_trunc						
age1	.0068285	.001723	3.96	0.000	.0034514	.0102056
male1	−.3500236	.058874	−5.95	0.000	−.4654146	−.2346326
lhincome1	.0175752	.0402519	0.44	0.662	−.0613171	.0964676
lsahbadvbad1	.1685531	.0555287	3.04	0.002	.0597187	.2773874
_cons	−.1156125	.3876709	−0.30	0.766	−.8754335	.6442084
alfa2						
_cons	.2188073	.0597075	3.66	0.000	.1017827	.3358318
/pi	.3470579	.0116303	29.84	0.000	.324627	.3701892

Stata also displays the list of constraints imposed, not shown here.

We compare the nested models LCNB2-Pan and LCH-Pan according to information criteria, displaying the results for the new model together with the ones stored earlier:

• estimates store lchurdle_pan
• estimates stats lcnb2_pan lchurdle_pan

The LCH-Pan outperforms the LCNB2-Pan, even penalizing for the larger number of parameters. Recall that these criteria are usually considered in the choice of the number of latent classes. We therefore compare the AIC and the BIC of the LCH-Pan with those of the (degenerate) one-class hurdle model to assess whether moving from one class to two classes improves the AIC and the BIC. We saw above that aic_hurdle = 84573.981 and bic_hurdle=84666.145, which are considerably smaller than the ones for the LCH-Pan shown in Table 11.18, providing evidence of unobserved time-invariant heterogeneity within the hurdle framework.

Table 11.18 **AIC** and **BIC** of **LCNB2-Pan and LCH-Pan** (with two latent classes) for the number of specialist visits

Model	nobs	11 (null)	11(model)	df	AIC	BIC
lcnb2 pan	8041		−41061.18	13	82148..36	82239.26
lchurdle pan	8041		−40674.74	23	81395..48	81556.3

The two specifications can also be compared by testing the restrictions that the parameters of the LCH-Pan are equal in the binary and the truncated parts (which corresponds to an LCNB2-Pan). We test those restrictions for each class:

• test [xb1_prob=xb1_trunc]
• test [xb2_prob=xb2_trunc]

For each class, the restricted NB2 specification is clearly rejected against the hurdle:

```
(1) [xb1_prob]age1−[xb1_trunc]age1=0
(2) [xb1_prob]male1−[xb1_trunc]male1=0
(3) [xb1_prob]lhincome1−[xb1_trunc]lhincome1=0
(4) [xb1_prob]lsahbad1−[xb1_trunc]lsahbad1–0
          chi2(4)=     226.12
    Prob > chi2=          0.0000
(1) [xb2_prob]age1−[xb2_trunc]age1–0
(2) [xb2_prob]male1−[xb2_trunc]male1–0
(3) [xb2_prob]lhincome1−[xb2_trunc]lhincome1=0
(4) [xb2_prob]lsahbad1−[xb2_trunc]lsahbad1–0
          chi2(4)=     166.51
    Prob>chi2=          0.0000
```

The latent class models estimated so far have assumed constant class memberships (π and $1-\pi$), following the most common approach in latent class models for health-care utilization. In the context of panel data models, this is similar to a random effects or random parameters specification that assumes no correlation between individual heterogeneity and the regressors. A generalization is obtained when individual heterogeneity is parameterized as a function of time-invariant individual characteristics z_i, as in Mundlak (1978). To implement this approach in the case of the latent class model, class membership can be modelled as a multinomial logit (as in, for example, Clark and Etilé 2003; Clark *et al.* 2005; Bago d'Uva 2005):

$$\pi_{ij} = \frac{\exp(z_i \gamma_j)}{\sum_{g=1}^{C} \exp(z_i \gamma_g)}, \quad j = 1, \ldots, C,$$

with $\gamma_c=0$. This uncovers the determinants of class membership. In a panel data context, this parameterization provides a way to account for the possibility that the observed regressors may be correlated with the individual effect. Let $z_i = \bar{x}_i$ be the average over the observed panel of the observations on the covariates. This is in line with what has been done in recent studies to allow for the correlation between covariates and random effects, following the suggestion of authors such as Mundlak (1978). The vectors of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c, \gamma_1, \ldots, \gamma_{c\_1}$ are estimated jointly by maximum likelihood.

In order to specify class membership probabilities as functions of $z_i = \bar{x}_i$, we create means of the covariates across the panel and the respective list:

```
• foreach var in $xvar{
  egen mean 'var'=rmean( 'var' 1 'var' 2 'var' 3 'var' 4)
  }
  • global xvarmean "meanage meanmale meanlhincome
  meanlsahbad"
```

A possible set of starting values for this model is the set of estimates of the LC Hurdle with constant class membership probabilities blchurdle. In vector γ (in π_i=exp(γ_iz)/(1+exp(γ_iz)), the coefficients of the covariates are initialized as zeros, except for the constant term, which starts at the estimate in the model with constant π. Starting values for γ are defined in vector initpi.

- scalar initpi0=blchurdle [1,2*k+2* (k+1)+ 1]
- mat initpi=blogit−blogit
- mat initpi= (initpi [1,1..k−1], initpi0)
- mat initlchurdle= (blchurdle [1,1..2*k+2* (k+1)],
  initpi,
  blchurdle [1,2*k+2* (k+1) +2..colsof(blchurdle)])

Estimation uses again the program lchurdle_pan, except that now the means of the covariates within individual, z_i, are included in the equation that corresponds to π (pi: $xvarmean). Estimates of the relevant parameters are shown:

- ml model If lchurdle_pan
  (xb1_prob: $xvar1) (xb1_trunc: $xvar1) (alfa1:)
  (xb2_prob: $xvar1) (xb2_trunc: $xvar1) (alfa2:)
  (pi:$xvarmean)
  (xb1_prob_w2:$xvar2) (xb1_prob_w3:$xvar3)
  (xb1_prob_w4:$xvar4)
  (xb2_prob_w2:$xvar2) (xb2_prob_w3:$xvar3)
  (xb2_prob_w4: $xvar4)
  (xb1_trunc_w2:$xvar2) (alfa1_w2:)
  (xb1_trunc_w3:$xvar3) (alfa1_w3:)
  (xb1_trunc_w4:$xvar4) (alfa1_w4:)
  (xb2_trunc_w2:$xvar2) (alfa2_w2:)
  (xb2_trunc_w3:$xvar3) (alfa2_w3:)
  (xb2_trunc_w4:$xvar4) (alfa2_w4:), technique(bfgs)
  const(1−$i)
- ml init initlchurdle, skip copy
- ml maximize, nooutput
- ml display, neq(7)

Table 11.19 shows the results displayed after estimation. The results under pi correspond to the logit model for the probability of belonging to class 1, within the LC hurdle for specialist visits. All variables are significant, especially meanlsahbad1, which is positively associated with that probability. Income also has a positive effect on the probability of belonging to class 1, while the association with male and age is negative. Since class membership is time invariant in this model and the covariates considered are averages across the panel, the estimated coefficients should be seen as a long-term association with class membership probabilities, unlike the effects on the distribution of the number of visits, conditional on the latent class to which the individual belongs, which represent short-term effects. Except for age, the estimated coefficients of the

hurdle model conditional on the latent class (xbj_prob, xbj_trunc and alfaj, for classes j=1,2) are substantially different from those in the model with constant class memberships (Table 11.17). This means that, in the restricted model, the coefficients of the conditional densities were also capturing the long-term effects that are disentangled in the specification that allows the class membership to be associated with the regressors. The estimated effects of lsahbad and lhincome become smaller throughout. The negative effects of male decrease in absolute value (in the second part for class 1, the effect was insignificantly negative and becomes insignificantly positive).

Table 11.19 **LCH-Pan** for the number of specialist visits (with two latent classes), with variable class membership

					Number of obs=	8041
					Wald chi2 (0)=	.
Log likelihood =−40498.679					Prob>chi2=	.
	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
xb1_prob						
age1	.0157471	.0024971	6.31	0.000	.010853	.0206413
male1	−.5021962	.0840317	−5.98	0.000	−.6668952	−.3374972
lhincome1	.3429209	.0414057	8.28	0.000	.2617671	.4240746
lsahbad1	.523134	.0719304	7.27	0.000	.382153	.6641149
_cons	−2.778549	.3838715	−7.24	0.000	−3.530923	−2.026174
xb1_trunc						
age1	.0029462	.0012085	2.44	0.015	.0005776	.0053148
male1	.0560945	.0418872	1.34	0.181	−.0260028	.1381918
lhincome1	.0446202	.0245924	1.81	0.070	−.00358	.0928203
lsahbad1	.4554045	.0385214	11.82	0.000	.379904	.5309049
_cons	.2219806	.2256706	0.98	0.325	−.2203256	.6642868
alfa1						
_cons	1.589425	.0925376	17.18	0.000	1.408055	1.770796
xb2_prob						
age1	.020847	.0017624	11.83	0.000	.0173927	.0243013
male1	−.4017409	.0697667	−5.76	0.000	−.5384811	−.2650007
lhincome1	.4398067	.0496106	8.87	0.000	.3425717	.5370417
lsahbad1	.1541941	.0630062	2.45	0.014	.0307043	.277684
_cons	−6.185499	.4539101	−13.63	0.000	−7.075146	−5.295851
xb2_trunc						
age1	.0067403	.0017525	3.85	0.000	.0033055	.0101751
male1	−.1621681	.0566564	−2.86	0.004	−.2732126	−.0511235
lhincome1	−.0654117	.0402197	−1.63	0.104	−.1442408	.0134175
lsahbad1	.0618263	.0587844	1.05	0.293	−.0533889	.1770415
_cons	.5546329	.3834023	1.45	0.148	−.1968217	1.306088
alfa2						
_cons	.2626443	.0725253	3.62	0.000	.1204975	.4047912

pi						
meanage	−.0191769	.0030117	−6.37	0.000	−.0250796	−.0132741
meanmale	−.5827018	.1081307	−5.39	0.000	−.7946341	−.3707694
meanlhincome	.5913606	.0785004	7.53	0.000	.4375027	.7452185
meanlsahbad	1.88748	.11945	15.80	0.000	1.653362	2.121597
_cons	−5.144374	.7123815	−7.22	0.000	−6.540616	−3.748131

Stata also displays the list of constraints imposed, not shown here.

Predictions for the individual probability of belonging to class 1 are computed and summarized, returning an average of 0.316 (Table 11.20):

- predict xbpi, eq(pi)
- gen pi=exp(xbpi) / (1+exp (xbpi))
- sum pi
- drop pi

Table 11.20 Summary statistics for individual π in LCH-Pan, with variable class membership

Variable	Obs	Mean	Std. Dev.	Min	Max
pi	8041	.3163629	.1449883	.0233796	.8481686

The computation of fitted values for each class requires the prediction of the linear indices $x_{it}\beta_{j1}$ and $x_{it}\beta_{j2}$ for each wave:

```
• foreach part in prob trunc{
      predict xb1 ‘part’_1, eq (xb1_‘part’)
      predict xb2 ‘part’_1, eq (xb2_‘part’)
      foreach wave in 2 3 4 {
          predict xb1 ‘part’_‘wave’, eq (xb1_‘part’ w ‘wave’)
          predict xb2 ‘part’_‘wave’, eq (xb2_‘part’ w ‘wave’)
      }
  }
```

We reshape the dataset back to long form and predict the probabilities of having at least one visit and expected number of visits, given that it is positive:

```
• reshape long y $xvar xb1prob_ xb2prob_
        xb1trunc_xb2trunc_, i (pidc) j (wave)
  • gen prob_c1=exp (xb1prob_) / (1+exp (xb1prob_))
  • gen prob_c2=exp (xb2prob_) / (1+exp (xb2prob_))
  • drop xb1prob_xb2prob_
  • predict a1, eq(alfa1)
  • predict a2, eq(alfa2)
  • gen pos_c1=exp(xb1trunc_)
  / (1−exp (−1/a1*log (a1*exp (xb1trunc_) +1)))
```

```
• gen pos_c2=exp (xb2trunc_)
/ (1−exp(−1/a2*log (a2*exp (xb2trunc_) +1)))
• drop xb1trunc_xb2trunc_a1 a2
```

For each class, the expected total number of visits is obtained as the product of the predictions for the binary and truncated parts, and summary statistics are computed:

```
• gen y_c1=prob_c1*pos_c1
• gen y_c2=prob_c2*pos_c2
• sumprob_c1 pos_c1 y_c1 prob_c2 pos_c2 y_c2
• drop prob_c1 pos_c1 y_c1 prob_c2 pos_c2 y_c2
```

The sample averages of predicted utilization conditional on the latent class, and decomposed into the probability of visiting a specialist at least once and the conditional number of visits, are shown in Table 11.21. The relative differences between latent classes are evident, being larger for the probability of visiting a specialist than for the conditional number of visits. The class of high users, class 1, is predicted to have an average total number of specialist visits that is more than seven times larger than the one for the class of low users. Looking again at Table 11.19, we see that longer-term poor health and higher incomes are associated with the probability of being a high user, while older individuals and males are more likely to be low users.

Table 11.21 Summary statistics of fitted values by latent class in LCH-Pan, with variable class membership

Variable	Obs	Mean	Std. Dev.	Min	Max
prob_c1	32164	.6846693	.1074067	2110298	.9353979
pos_c1	32164	3.906074	.6620348	2.914752	5.896463
Y_c1	32164	2.719638	.8263821	.6501456	5.333827
prob_c2	32164	.1886456	.0755504	.0170471	.613218
pos_c2	32164	1.952128	.1860879	1.576299	2.998852
Y_c2	32164	.3769873	.1768242	.0330797	1.311424

We test for the equality of coefficients across classes and conclude that there are significant differences both in the binary and in the truncated parts:

```
• test [xb1_prob=xb2_prob]
• test [xb1_trunc=xb2_trunc]

(1) [xb1_prob]age1−[xb2_prob]age1=0
   (2) [xb1_prob]male1−[xb2_prob]male1=0
   (3) [xb1_prob]lhincome1−[xb2_prob]lhincome1=0
   (4) [xb1_prob]lsahbad1−[xb2_prob]lsahbad1=0
         chi2(4)=    23.46
    Prob > chi2=     0.0001
```

```
(1) [xb1_trunc]age1−[xb2_trunc]age1=0
(2) [xb1_trunc]male1−[xb2_trunc]male1–0
(3) [xb1_trunc]lhincome1−[xb2_trunc]lhincome1=0
(4) [xb1_trunc]lsahbad1−[xb2_trunc]lsahbad1=0
          chi2(4)=     40.75
      Prob>chi2=     0.0000
```

Individual tests of equality of parameters across classes, coupled with the results in Table 11.19, show us that the effects of poor health are significantly larger for high users, in both the binary and the truncated parts, while the effect of income in the binary part is larger for low users (difference significant at 10%), and in the truncated part it is larger, in absolute value, for low users:

```
• test [xb1_prob]lhincome1= [xb2_prob]lhincome1
  • test [xb1_trunc]lhincome1= [xb2_trunc]lhincome1
  • test [xb1_prob]lsahbad1= [xb2_prob]lsahbad1
  • test [xb1_trunc]lsahbad1= [xb2_trunc]lsahbad1

(1) [xb1_prob]lhincome1=[xb2_prob]lhincome1=0
          chi2(1)=     2.89
      Prob > chi2=     0.0890
  (1) [xb1_trunc]lhincome1 − [xb2_trunc]lhincome1=0
          chi2 (1)=     5.41
      Prob > chi2=     0.0200
  (1) [xb1_prob]lsahbad1 − [xb2_prob]lsahbad1=0
          chi2(1)=     16.06
      Prob > chi2=     0.0001
  (1) [xb1_trunc]lsahbad1 − [xb2_trunc]lsahbad1=0
          chi2(1)=     30.25
      Prob > chi2=     0.0000
```

Similar tests for ‘age’ and ‘male’ show that the effect of age on the probability of visiting a specialist is significantly higher for low users and that the effect of male on the conditional positive number of visits is significantly larger, in absolute value, for low users.

Information criteria are displayed for the LCH-Pan and the restricted versions estimated above:

```
• estimates store lchurdle_pan_varpi
• estimates stats lcnb2_pan lchurdle_pan
lchurdle_pan_varpi
```

The more general specification, the latent class hurdle model with class membership probabilities modelled as functions of the covariates, is the preferred specification according to the information criteria (Table 11.22).

Table 11.22 AIC and BIC of LCNB2-Pan and LCH-Pan (with two latent classes) with constant and variable class memberships

Model	nobs	ll (null)	ll(model)	df	AIC	BIC
lcnb2_pan	8041	.	−41061.18	13	82148.36	82239.26
lchurdle_pan	8041	.	−40674.74	23	81395.48	81556.3
lchurdle_p~i	8041	.	−40498.68	27	81051.36	81240.15

The latent class panel data model accounts for the panel features of the data in a flexible way that assumes no distribution for the unobserved individual effects. It can also be seen as a discrete approximation of an underlying continuous mixing distribution (Heckman and Singer 1984). The number of points of support needed for the finite mixture model is low, usually two or three. The specification used here allows for correlation between latent heterogeneity and the covariates. The conventional fixed effects models that have been developed for binary dependent variables (conditional logit) and for counts (fixed effects Poisson and NegBin) also offer a distribution-free approach to the individual heterogeneity that is robust to correlation between covariates and individual effects. However, although fixed effects models account for intercept heterogeneity, they do not accommodate different responses to the covariates across individuals, while the latent class model accommodates both intercept heterogeneity and slope heterogeneity. Furthermore, fixed effects models do not allow the estimation of the effects of time-invariant regressors. In these models, the coefficients of time-invariant regressors are absorbed into the intercept.

Bibliography

Adams, P., Hurd, M.D., McFadden, D., Merrill, A. and Ribeiro, T. (2003) ‘‘Healthy, wealthy and wise? Tests for direct causal paths between health and socioeconomic status’, *Journal of Econometrics,* 112:3–56.

Amemiya, T. and MaCurdy, T. (1986) ‘Instrumental variable estimation of an error-components model’, *Econometrica,* 54:869–880.

Anderson, K.H. and Burkhauser, R.V. (1985) ‘The Retirement-Health Nexus: A New Measure of an Old Puzzle’, *Journal of Human Resources,* 20:315–330.

Atella, V., Brindisi, F., Deb, P. and Rosati, F.C. (2004) ‘Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach’, *Health Economics,* 13:657–668.

Au, D., Crossley, T.F. and Schellhorn, M. (2005) ‘The effects of health shocks and long-term health on the work activity of older Canadians’, *Health Economics,* 14: 999–1018.

Auster, R., Levenson, I. and Sarachek, D. (1969) ‘The production of health: an exploratory study’, *Journal of Human Resources,* 4:411–436.

Bago d’Uva, T. (2005) ‘Latent class models for use of primary care: evidence from a British panel’, *Health Economics,* 14:873–892.

Bago d’Uva, T. (2006) ‘Latent class models for health care utilisation’, *Health Economics,* 15:329–343.

Bago d’Uva, T., van Doorslaer, E., Lindeboom, M., O’Donnell, O. and Chatterji, S. (2006) ‘Does reporting heterogeneity bias the measurement of health disparities?’, HEDG Working Paper 06/03, University of York.

Balia, S. and Jones, A.M. (2005) ‘Mortality, Lifestyle and Socio-Economic Status’, HEDG Working Paper 05/02, University of York.

Baltagi, B.H. (2005) *Econometric Analysis of Panel Data,* Chichester: John Wiley & Sons.

Baltagi, B.H. and Khanti-Akom, S. (1990) ‘On efficient estimation with panel data: An empirical comparison of instrumental variables estimators’, *Journal of Applied Econometrics,* 5:401–406.

Bazzoli, G.J. (1985) ‘The early retirement decision: new empirical evidence on the influence of health’, *Journal of Human Resources,* 20(2): 214–234.

Belloc, N.B. (1973) ‘Relationship of health practices and mortality’, *Preventive Medicine,* 2:67–81.

Belloc, N.B. and Breslow, L. (1972) ‘Relationship of physical health status and health practices’, *Preventive Medicine,* 1:409–421.

Benzeval, M., Taylor, J. and Judge, K. (2000) ‘Evidence on the relationship between low income and poor health: Is the Government doing enough?’, *Fiscal Studies,* 21:375–399.

Blundell, R., Meghir, C. and Smith, S. (2002) ‘Pension incentives and the pattern of early retirement’, *Economic Journal,* 112:153–170.

Bound, J. (1991) ‘Self reported versus objective measures of health in retirement models’, *Journal of Human Resources,* 26:107–137.

Bound, J., Schoenbaum, M., Stinebrickner, T.R. and Waidmann, T. (1999) ‘The dynamic effects of health on the labor force transitions of older workers’, *Labour Economics,* 6:179–202.

Breusch, T., Mizon, G.E. and Schmidt, P. (1989) ‘Efficient estimation using panel data’, *Econometrica,* 57:695–700.

Burström, B. and Fredlund, P. (2001) ‘Self rated health: is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes?’, *Journal of Epidemiology and Community Health,* 55:836–840.

Butler, J. and Moffitt, R. (1982) 'A computationally efficient quadrature procedure for the one-factor multinomial probit model', *Econometrica,* 50:761–764.
Butler, J.S., Burkhauser, R.V., Mitchel, J.M. and Pincus, T.P. (1987) 'Measurement error in self-reported health variables', *The Review of Economics and Statistics,* 69: 644–650.
Cameron, A.C. and Trivedi, P.K. (1998) *Regression Analysis for Count Data,* Cambridge: Cambridge University Press.
Cameron, A.C. and Trivedi, P.K. (2005) *Microeconometrics. Methods and Applications,* Cambridge: Cambridge University Press.
Cameron, A.C., Trivedi, P.K., Milne, F. and Piggot, J.R. (1988) 'A microeconometric model of the demand for health care and health insurance in Australia', *Review of Economic Studies,* 55:85–106.
Cappellari, L. and Jenkins, S.P. (2003) 'Multivariate probit regression using simulated maximum likelihood', *The Stata Journal,* 3:278–294.
Chamberlain, G. (1980) 'Analysis of covariance with qualitative data', *Review of Economic Studies,* 47:225–238.
Chamberlain, G. (1984) 'Panel data', in Griliches, Z. and Intriligator, M.D. (eds), *Handbook of Econometrics Volume 1,* Amsterdam: Elsevier.
Cheung, Y.B. (2000) 'Marital status and mortality in British women: a longitudinal study', *International Journal of Epidemiology,* 29:93–99.
Clark, A. and Etilé, F. (2006) 'Don't give up on me baby: Spousal correlation in smoking behaviour', *Health Economics,* 25:958–978.
Clark, A., Etilé, F., Postel-Vinay, F., Senik, C. and Van der Straeten, K. (2005) 'Heterogeneity in reported well-being: Evidence from!twelve European countries', *The Economic Journal,* 115:118–132.
Contoyannis, P. and Jones, A.M. (2004) 'Socioeconomic status, health and lifestyle', *Journal of Health Economics,* 23:965–995.
Contoyannis, P., Jones, A.M. and Leon-Gonzalez, R. (2004) 'Using simulationbased inference with panel data in health economics', *Health Economics,* 13: 101–122.
Contoyannis, P., Jones, A.M. and Rice, N. (2004) 'The dynamics of health in the British Household Panel Survey', *Journal of Applied Econometrics,* 19:473–503.
Contoyannis, P. and Rice, N. (2001) 'The impact of health on wages: Evidence from the British Household Panel Survey', *Empirical Economics,* 26:599–622.
Cornwell, C. and Rupert, P. (1988) 'Efficient estimation with panel data: An empirical comparison of instrumental variables estimators', *Journal of Applied Econometrics,* 3:149–155.
Cox, B.D., Blaxter, M., Buckle, A.L.J., Fenner, N.P., Golding, J.F., Gore, M., Huppert, F.A., Nickson, J., Roth, M., Stark, J., Wadsworth, M.E.J. and Whichelow, M. (1987) *The Health and Lifestyle Survey,* London: Health Promotion Research Trust.
Cox, B.D., Huppert, F.A. and Wichelow, M.J. (1993) *The Health and Lifestyle Survey: seven years on,* Aldershot: Dartmouth.
Crossley, T.F. and Kennedy, S. (2002) 'The reliability of self-assessed health status', *Journal of Health Economics,* 21:643–658.
Currie, J. and Madrian, B.C. (1999) 'Health, health insurance and the labor market', in Ashenfelter, O. and Card, D. (eds) *Handbook of Labour Economics Volume 3,* Amsterdam: Elsevier.
Davidson, R. and Mackinnon, J.G. (1993) *Estimation and Inference in Econometrics,* Oxford: Oxford University Press.
Deaton, A.S. (1997) *The Analysis of Household Data: A Microeconometric approach to Development Policy,* Baltimore: Johns Hopkins Press.
Deaton, A.S. and Paxson, C.H. (1998) 'Ageing and inequality in income and health', *American Economic Review, Papers and Proceedings,* 88:248–253.
Deb, P. (2001) 'A discrete random effects probit model with application to the demand for preventive care', *Health Economics,* 10:371–383.

Deb, P. and Holmes, A.M. (2000) 'Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models', *Health Economics,* 9: 475–489.
Deb, P. and Trivedi, P.K. (1997) 'Demand for medical care by the elderly: a finite mixture approach', *Journal of Applied Econometrics,* 12:313–336.
Deb, P. and Trivedi, P.K. (2002) 'The structure of demand for health care: latent class versus two-part models', *Journal of Health Economics,* 21:601–625.
Disney, R., Emmerson, C. and Wakefield, M. (2006) 'Ill-health and retirement in Britain: a panel data-based analysis', *Journal of Health Economics,* 25:621–649.
Disney, R. and Gosling, A. (1998) 'Does it pay to work in the public sector?', *Fiscal Studies,* 19:347–374.
Ettner, S. (1996) 'New evidence on the relationship between income and health', *Journal of Health Economics,* 15:67–85.
Feeny, D., Furlong, W., Boyle, M. and Torrance, G. (1995) 'Multi-attribute health status classification systems: Health Utilities Index', *Pharmacoeconomics,* 7:490–502.
Fitzgerald, J., Gottshalk, P. and Moffitt, R. (1998) 'An analysis of sample attrition in panel data. The Michigan Panel Study on Income Dynamics', *Journal of Human Resources,* 33:251–299.
Forster, M. and Jones, A.M. (2001) 'The role of tobacco taxes in starting and quitting smoking: duration analysis of British data', *Journal of the Royal Statistical Society Series A,* 164:517–547.
Frijters, P., Haisken-DeNew, J.P. and Shields, M.A. (2003) 'Estimating the causal effect of income on health: evidence from post reunification East Germany', Centre for Economic Policy Discussion Paper No. 465, Australian National University.
Gerdtham, U.G. (1997) 'Equity in health care utilization: further tests based on hurdle models and Swedishk micro data', *Health Economics,* 6:303–319.
Greene, W. (2001) 'Fixed and Random Effects in Nonlinear Models', Working paper 01–01, New York University, Department of Economics, Stern School of Economics.
Greene, W.H. (2002) *LIMDEP, Version 8.0,* New York: Econometric Software.
Greene, W.H. (2003) *Econometric Analysis,* 5th edition, New York: Macmillan.
Griliches, Z. (1977) 'Estimating the return to schooling: Some econometric problems', *Econometrica,* 45:1–22.
Groot, W. (2000) 'Adaptation and scale of reference bias in self-assessments of quality of life', *Journal of Health Economics,* 19:403–420.
Grootendorst, P. (1995) 'A comparison of alternative models of prescription drug utilization', *Health Economics,* 4:183–198.
Grootendorst, P., Feeny, D. and Furlong, W. (1997) 'Does it matter whom and how you ask? Inter and intra-rater agreement in the Ontario health survey', *Journal of Clinical Epidemiology,* 50:127–136.
Grossman, M. and Joyce, T.J. (1990) 'Unobservables, pregnancy resolutions, and birth weight production functions in New York City', *Journal of Political Economy,* 98:983–1007.
Gurmu, S. (1997) 'Semi-parametric estimation of hurdle regression models with an application to medicaid utilizations', *Journal of Applied Econometrics,* 12:225–242.
Hakkinen, U., Rosenqvist, G. and Aro, S. (1996) 'Economic Depression and the use of physician services in Finland', *Health Economics,* 5:421–434.
Harkness, S. (1996) 'The gender earnings gap: evidence from the UK', *Fiscal Studies,* 17:1–36.
Hausman, J. (1978) 'Specification tests in econometrics', *Econometrica,* 46:1251–1271.
Hausman, J. and Taylor, W. (1981) 'Panel data and unobservable individual effects', *Econometrica,* 49:1377–1398.
Hausman, J. and Wise, D. (1979) 'Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment', *Econometrica,* 47:455–474.
Heckman, J.J. (1976) 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *The Annals of Economic and Social Measurement,* 5:475–492.

Heckman, J. (1981) 'The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process', in Manski, C.F. and McFadden, D. (eds), *Structural Analysis of Discrete Data with econometric applications,* Cambridge, MA: MIT Press.

Heckman, J. and Singer, B. (1984) 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica,* 52: 271–320.

Hernández-Quevedo, C., Jones, A.M. and Rice, N. (2004) 'Reporting bias and heterogeneity in self-assessed health. Evidence from the British Household Panel Survey', ECuity III Working Paper #19, Erasmus University.

Hildreth, A. (1999) 'What has happened to the union wage differential in Britain in the 1990s?', *Oxford Bulletin of Economics and Statistics,* 61:5–31.

Horowitz, J.L. and Manski, C.F. (1998) 'Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations', *Journal of Econometrics,* 84:37–58.

Humphrey, A., Costigan, P., Pickering, K., Stratford, N. and Barnes, M. (2003) 'Factors affecting the labour market participation of older workers', Research Report No 200. Department for Work and Pensions, London.

Idler, E.L. and Benyamini, Y. (1997) 'Self-rated health and mortality: a review of twenty-seven community studies', *Journal of Health and Social Behavior,* 38: 21–37.

Idler, E.L. and Kasl, S.V. (1995) 'Self-ratings of health: do they also predict change in functional ability?', *Journal of Gerontology,* 50B: S344-S353.

Jenkins, S.P. (1995) 'Easy Estimation Methods for Discrete-Time Duration Models', *Oxford Bulletin of Economics and Statistics,* 57:129–138.

Jenkins, S.P. (1997) 'Discrete time proportional hazards regression', STATA Technical Bulletin. STB-39; pp. 22–23.

Jenkins, S.P. (2004) *Survival Analysis.* Unpublished manuscript, Institute for Social and Economic Research, University of Essex.

Jimenez-Martin, S., Labeaga, J.M. and Martinez-Granado, M. (2002) 'Latent class versus two-part models in the demand for physician services across the European Union', *Health Economics,* 11:301–321.

Jones, A.M. (2000) 'Health Econometrics', in Newhouse, J.P. and Culyer, A.J. (eds), *Handbook of Health Economics,* Amsterdam: Elsevier.

Jones, A.M., Koolman, X. and Rice, N. (2006) 'Health-related non-response in the BHPS and ECHP: using inverse probability weighted estimators in nonlinear models', *Journal of the Royal Statistical Society Series A,* 169:543–569.

Jones, A.M. and O'Donnell, O.A. (2002) *Econometric Analysis of Health Data,* Chichester: John Wiley & Sons.

Kapteyn, A., Smith, J. and van Soest, A. (2004) 'Self-reported work disability in the US and the Netherlands', RAND Working Paper. Santa Monica.

Kenkel, D. (1995) 'Should you eat breakfast? Estimates from health production functions', *Health Economics,* 4:15–29.

Kerkhofs, M.J.M. and Lindeboom, M. (1995) 'Subjective health measures and state dependent reporting errors', *Health Economics,* 4:221–235.

King, G., Murray, C.J.L., Salomon, J. and Tandon, A. (2004) 'Enhancing the validity and cross-cultural comparability of measurement in survey research', *American Political Science Review,* 98:184–191.

Knapp, L.G. and Seaks, T.G. (1998) 'A Hausman test for a dummy variable in probit', *Applied Economics Letters,* 5:321–323.

Kunst, A., Giskes, K. and Mackenbach, J. (2004) 'Socioeconomic inequalities in smoking in the European Union. Applying an equity lens to tobacco control policy', Department of Public Health. Erasmus Medical Center, Rotterdam.

Lazear, E.P. (1986) 'Retirement from the labour force', in Ashenfelter, O.C. and Layard, R. (eds), *Handbook of Labour Economics Volume 1,* Amsterdam: Elsevier.

Lindeboom, M. (2006) 'Health and work of older workers', in Jones, A.M. (ed.) *Elgar Companion to Health Economics,* Cheltenham: Edward Elgar.

Lindeboom, M. and van Doorslaer, E. (2004) 'Cut-point shift and index shift in self-reported health', *Journal of Health Economics,* 23:1083–1099.

Little, J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data,* New York: John Wiley and Sons.

Lumsdaine, R.L. and Mitchell, O.S. (1999) 'New developments in the economic analysis of retirement', in Ashenfelter, O.C. and Card, D. (eds), *Handbook of Labour Economics Volume 3,* Amsterdam: Elsevier.

McGinnis, J.M. and Foege, W.H. (1993) 'Actual causes of death in the United States', *The Journal of the American Medical Association,* 270:2207–2212.

Maddala, G.S. (1983) *Limited-dependent and Qualitative Variables in Econometrics,* Cambridge: Cambridge Unversity Press.

Meyer, B.D. (1990) 'Unemployment insurance and unemployment spells', *Econometrica,* 58:757–782.

Michaud, P.C. (2003) 'Joint labour supply dynamics of older couples', Centre for Economic Research, Discussion Paper: 69, Tilburg University.

Mincer, J. (1974) *Schooling, Experience and Earnings,* NBER, New York: Columbia University Press.

Moffitt, R., Fitzgerald, J. and Gottschalk, P. (1999) 'Sample attrition in panel data: the role of selection observables', *Annales d'Economie et de Statistique,* 55–56: 129–152.

Mokdad, A.L., Marks, J.S., Stroup, D.F. and Gerberding, J.L. (2004) 'Actual causes of death in the United States, 2000', *The Journal of the American Medical Association,* 291:1238–1245.

Mullahy, J. (1986) 'Specification and testing in some modified count data models', *Journal of Econometrics,* 33:341–365.

Mullahy, J. and Portney, P. (1990) 'Air pollution, cigarette smoking, and the production of respiratory health', *Journal of Health Economics,* 9:193–205.

Mullahy, J. and Sindelar, J. (1996) 'Employment, unemployment, and problem drinking', *Journal of Health Economics,* 15:409–434.

Mundlak, Y. (1978) 'On the pooling of time series and cross-section data', *Econometrica,* 46:69–85.

Murray, C.J.L., Tandon, A., Salomon, J. and Mathers, C.D. (2001) 'Enhancing cross-population comparability of survey results', GPE Discussion Paper Nr 35, WHO/EIP, Geneva.

Murray, C.J.L., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R. and Chatterji, S. (2003) 'Empirical evaluation of the anchoring vignettes approach in health surveys', in Murray, C.J.L. and Evans, D.B. (eds), *Health Systems Performance Assessment:Debates, Methods and Empiricism,* Geneva: World Health Organization.

Nagin, D. and Land, K. (1993) 'Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric mixed Poisson model', *Criminology,* 31:327–362.

Narendranathan, W. and Stewart, M.B. (1993) 'How does the benefit effect vary as unemployment spells lengthen?', *Journal of Applied Econometrics,* 8:361–381.

Nicoletti, C. (2002) 'Non-response in dynamic panel data models', Working papers of the Institute for Social and Economic Research, paper 2002–31, Colchester: University of Essex.

Nicoletti, C. and Peracchi, F. (2005) 'A cross-country comparison of survey nonparticipation in the ECHP', *Journal of the Royal Statistical Society Series A,* 168:763–781.

Peracchi, F. (2002) 'The European Community Household Panel: a review', *Empirical Economics,* 27:63–90.

Peto, R., Lopez, A.D., Boreham, J. and Thun, M. (2005) *Mortality from Smoking in Developed Countries 1950–2000,* Oxford: Oxford University Press.

Pohlmeier, W. and Ulrich, V. (1995) 'An econometric model of the two-part decision making process in the demand for health care', *The Journal of Human Resources,* 30:339–361.

Prentice, R. and Gloeckler, L. (1978) 'Regression analysis of grouped survival data with applications to breast cancer data', *Biometrics,* 34:57–67.

Pudney, S. and Shields, M. (2000) 'Gender, race, pay and promotion in the British nursing profession: estimation of a generalized probit model', *Journal of Applied Econometrics,* 15:367–399.

Riphahn, R.T. (1997) 'Disability, retirement and unemployment: substitute pathways for labour force exit? An empirical test for the case of Germany', *Applied Economics,* 29:551–561.

Robins, J., Rotnitzky, A. and Zhao, L.P. (1995) 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association,* 90:106–121.

Rosenzweig, M.R. and Schultz, T.P. (1983) 'Estimating a household production function: heterogeneity, the demand for health inputs, and their effect on birth weight', *Journal of Political Economy,* 91:723–746.

Rotnitzky, A. and Robins, J. (1994) 'Analysis of semi-parametric regression models with non-ignorable non-response', *Statistics in Medicine,* 16:81–102.

Rubin, D.B. (1976) 'Inference and missing data', *Biometrika,* 63:581–592.

Sadana, R., Mathers, C.D., Lopez, A.D., Murray, C.J.L. and Iburg, K. (2000) 'Comparative analysis of more than 50 household surveys on health status', GPE Discussion Paper No 15, EIP/GPE/EBD, World Health Organization, Geneva.

Salas, C. (2002) 'On the empirical association between poor health and low socioeconomic status at old age', *Health Economics,* 11:207–220.

Salomon, J., Tandon, A., Murray, C.J.L. and WHSPSC Group (2004) 'Comparability of self-rated health: cross sectional multi-country survey using anchoring vignettes', *British Medical Journal,* 328:258.

Santos Silva, J.M.C. and Windmeijer, F. (2001) 'Two-part multiple spell models for health care demand', *Journal of Econometrics,* 104:67–89.

Smith, J.P. (1999) 'Healthy bodies and thick wallets: the dual relationship between health and economic status', *Journal of Economic Perspectives,* 13:145–166.

Stern, S. (1989) 'Measuring the effect of disability on labour force participation', *Journal of Human Resources,* 24:361–395.

Stewart, M. (2006) 'Redprob—A Stata program for the Heckman estimator of the random effects dynamic probit model', mimeo, University of Warwick.

Tambay, J.-L. and Catlin, G. (1995) 'Sample design of the National Population Health Survey', *Health Reports,* 7:29–38.

Tandon, A., Murray, C.J.L., Salomon, J.A. and King, G. (2003) 'Statistical models for enhancing cross-population comparability. Health systems performance assessment: debates, methods and empiricisms', in Murray, C.J.L. and Evans, D.B. (eds), *Health Systems Performance Assessment: Debates, Methods and empiricism,* Geneva: World Health Organization.

Taylor, M., Brice, J., Buck, N. and Prentice, E. (1998) *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* Colchester: University of Essex.

Terza, J.V. (1985) 'Ordinal probit: a generalization', *Communications in Statistics,* 14:1–11.

Torrance, G.W., Furlong, W., Feeny, D. and Boyle, M. (1995) 'Multi-attribute preference functions: Health Utilities Index', *Pharmacoeconomics,* 7:503–520.

Torrance, G.W., Feeny, D., Furlong, W.J., Barr, R.D., Zhang, Y. and Wang, Q. (1996) 'Multiattribute utility function for a comprehensive health status classification system', *Medical Care,* 34:702–722.

Train, K.E. (2003) *Discrete choice methods with simulation,* Cambridge: Cambridge University Press.

Uebersax, J.S. (1999) 'Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models', *Applied Psychological Measurement,* 23:283–297.

Üstün, T.B., Chatterji, S., Villanueva, M., Benib, L., Celik, C., Sadana, R., Valentine, N.B., Ortiz, J.P., Tandon, A., Salomon, J., Yang, C., Xie, W.J., Ozaltin, E., Mathers, C.D. and Murray, C.J.L. (2003) 'WHO multi-country survey study on health and responsiveness 2001–2', in Murray, C.J.L. and Evans, D.B. (eds), *Health Systems Performance Assessment: Debates, Methods and Empiricism,* Geneva: World Health Organization.

van Doorslaer, E. and Gerdtham, U.-G. (2003) 'Does inequality in self-assessed health predict inequality in survival by income? Evidence from Swedish data', *Social Science and Medicine,* 57:1621–1629.

van Doorslaer, E. and Jones, A.M. (2003) 'Inequalities in self-reported health: valida-tion of a new approach to measurement', *Journal of Health Economics,* 22:61–87.

van Doorslaer, E. and Koolman, X. (2004) 'Explaining the differences in incomerelated health inequalities across European countries', *Health Economics,* 13:609–628.

van Doorslaer, E., Jones, A.M. and Koolman, X. (2004) 'Explaining income-related inequalities in doctor utilisation in Europe', *Health Economics,* 13:629–647.

van Doorslaer, E., Wagstaff, A., Bleichrodt, H. *et al.* (1997) 'Income-related inequalities in health: some international comparisons', *Journal of Health Economics,* 16:93–112.

van Doorslaer, E., Wagstaff, A., van der Burg. H., Christiansen, T., De Graeve, D., Duchesne, I. Gerdtham, U.G., Gerfin, M., Geurts, J., Gross, L., Hakkinen, U., John, J., Klavus, J., Leu, R.E., Nolan, B., O'Donnell, O., Propper, C., Puffer, F., Schellhorn, M., Sundberg, G. and Winkelhake, O. (2000) 'Equity in the delivery of health care in Europe and the US', *Journal of Health Economics,* 19:553–583.

van Ourti, T. (2003) 'Socioeconomic inequality in ill-health amongst the elderly. Should one use current income or permanent income?', *Journal of Health Economics,* 22:187–217.

Verbeek, M. and Nijman, T.E. (1992) 'Testing for selectivity bias in panel data models', *International Economic Review,* 33:681–703.

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y.T., Gupta, P.C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M.J., Wichmann, H.E., Wu, A.H., Zaridze, D., Peto, R. and Doll, R. (2004) 'Tobacco and cancer: recent epidemiological evidence', *Journal National Cancer Institute,* 96:99–106.

Wagstaff, A., Paci, P. and van Doorslaer, E. (1991) 'On the measurement of inequalities in health', *Social Science and Medicine,* 33:545–557.

Wagstaff, A. and van Doorslaer, E. (2000) 'Measuring and testing for inequity in the delivery of health care', *Journal of Human Resources,* 35:716–733.

Wagstaff, A., van Doorslaer, E. and Paci, P. (1989) 'Equity in the finance and delivery of health care: some tentative cross-country comparisons', *Oxford Review of Economic Policy,* 5:89–112.

Wang, P., Cockburn, I.M. and Puterman, M.L. (1998) 'Analysis of patent data a mixed Poisson regression model approach', *Journal of Business and Economic Statistics,* 16:27–41.

Wedel, M., DeSarbo, W.S., Bult, J.R. and Ramaswamy, V. (1993) 'A latent class Poisson regression model for heterogeneous count data', *Journal of Applied Econometrics,* 8:397–411.

Wilde, J. (2000) 'Identification of multiple equation probit models with endogenous dummy variables', *Economic Letters,* 69:309–312.

Wooldridge, J.M. (2002a) 'Inverse probability weighted M-estimators for sample selection, attrition and stratification', *Portuguese Economic Journal,* 1:117–139.

Wooldridge, J.M. (2002b) *Econometric analysis of cross-section and panel data,* Cambridge, MA: The MIT Press.

Wooldridge, J.M. (2005) 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', *Journal of Applied Econometrics,* 20:39–54.

Index