SECOND EDITION

# APPLIED ECONOMETRICS FOR HEALTH ECONOMISTS

## A PRACTICAL GUIDE

ANDREW JONES

OHE
Research
Office of Health Economics

# Applied Econometrics for Health Economists

## A practical guide

## Second Edition

Andrew Jones
*Professor of Economics*
*Department of Economics and Related Studies*
*University of York*

OHE
Research
Office of Health Economics

Radcliffe Publishing
Oxford • Seattle

**Radcliffe Publishing Ltd**
18 Marcham Road
Abingdon
Oxon OX14 1AA
United Kingdom

**www.radcliffe-oxford.com**
Electronic catalogue and worldwide online ordering facility.

First edition 2001 (published by the OHE)

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library.

ISBN-10: 1 84619 171 8
ISBN-13: 978 1 84619 171 8

Typeset by Lapiz Digital Services, Chennai, India
Printed and bound by Biddles Ltd, King's Lynn, Norfolk, UK

# Contents

# Preface

Given the extensive use of individual-level survey data in health economics, it is important to understand the econometric techniques available to applied researchers. Moreover, it is just as important to be aware of their limitations and pitfalls. The purpose of this book is to introduce readers to the appropriate econometric techniques for use with different forms of survey data – known collectively as microeconometrics. There is a strong emphasis on applied work, illustrating the use of relevant computer software applied to large-scale survey data sets. The aim is to illustrate the steps involved in doing microeconometric research:

- formulate empirical problems involving large survey data sets
- construct usable data sets and know the limitations of survey design
- select an appropriate econometric method
- be aware of the methods of estimation that are available for microeconometric models and the software that can be used to implement them
- interpret the results of the analysis and describe their implications in a statistically and economically meaningful way.

The standard linear regression model, familiar from econometric textbooks, is designed to deal with a dependent variable that varies continuously over a range between minus infinity and plus infinity. Unfortunately, this standard model is rarely applicable with survey data, where qualitative and categorical variables are more common. This book therefore deals with practical analysis of qualitative and categorical variables. The book assumes basic familiarity with the principles of statistical inference – estimation and hypothesis testing – and with the linear regression model. An accessible and clear overview of the linear regression model is given in the fifth edition of Peter Kennedy's book *A Guide to Econometrics*, published by MIT Press, and the material is covered in many other introductory econometrics textbooks.

Technical details or derivations are avoided in the main text and the book concentrates on the intuition behind the models and their interpretation. Key terms are marked in **bold** and defined in the Glossary. Formulae and more technical details are presented in the Technical appendix; the structure of the appendix follows that of the main text, with the numbered sections in the appendix corresponding to the chapters in the main text. References are kept to a minimum to maintain the flow of the text and are augmented with a list of further Recommended reading for readers who would like to pursue the topics in more detail. All of the results presented are estimated using Stata (www.stata.com/). Examples of relevant Stata commands are described and explained in an appendix to each chapter, and a separate Software appendix lists the full set of Stata commands that can be used to compute the methods and empirical examples used in the text. To give a feel for the way that the software package presents results, the tables are reproduced as they appear in the Stata output. The text refers to key results only and readers who want a full explanation of all of the statistics listed are encouraged to consult the Stata user manuals.

**Andrew Jones**
*November 2006*

# About the author

**Andrew Jones** is professor of economics at the University of York, where he directs the graduate programme in health economics, and visiting professor at the University of Bergen. He is research director of the Health, Econometrics and Data Group (HEDG) at the University of York. He researches and publishes extensively in the area of microeconometrics and health economics. He is an organiser of the European Workshops on Econometrics and Health Economics and coordinator of the Marie Curie Training Programme in Applied Health Economics. He has edited the *Elgar Companion to Health Economics*, is joint editor of *Health Economics* and of *Health Economics Letters*, and is an associate editor of the *Journal of Health Economics*.

# Acknowledgements

I am grateful to my colleagues in the Health, Econometrics and Data Group (HEDG) at the University of York for their helpful comments and suggestions on earlier versions of the book, and to Hugh Gravelle, Carol Propper and Frank Windmeijer for their insightful and comprehensive reviews of the material. Thanks to Jon Sussex, who provided me with the original challenge of preparing a non-technical guide to econometrics for health economists, 'without equations'.

Chapter 1

# Introduction: the evaluation problem and linear regression

## The evaluation problem

The evaluation problem is how to identify causal effects from empirical data. An understanding of the implications of the evaluation problem for statistical analysis will help to provide a motivation for many of the econometric methods discussed below.

Consider an outcome $y_{it}$, for individual i at time t; for example an individual's level of use of healthcare services over the past year. The problem is to identify the effect of a treatment, for example whether the individual has purchased private health insurance, on the outcome. The causal effect of interest is the difference between the outcome with the treatment and the outcome without the treatment. But this pure treatment effect cannot be identified from empirical data. This is because the counterfactual can never be observed. The basic problem is that the individual 'cannot be in two places at the same time'; that is, we cannot observe their use of healthcare at time t, both with and without the influence of insurance.

One response to this problem is to concentrate on the **average treatment effect** and attempt to estimate it with sample data by comparing the average outcome among those receiving the treatment with the average outcome among those who do not receive the treatment. The problem for statistical inference arises if there are unobserved factors that influence both whether an individual is selected into the treatment group and also how they respond to the treatment. This will lead to biased estimates of the treatment effect. For example, someone who knows they have a high risk of illness may be more prone to take out health insurance and they will also tend to use more healthcare. Unless the analyst is able to control for their level of risk, this will lead to spurious evidence of a positive relationship between having health insurance and using healthcare.

A randomised experimental design – which randomises the allocation of individuals into treatments – may be able to control for this bias and, in some circumstances, a natural experiment may mimic the features of a controlled experiment. However, the vast majority of econometric studies rely on observational data gathered in a non-experimental setting. In the absence of experimental data, attention has to focus on alternative estimation strategies.

- **Instrumental variables (IV)**: variables (or 'instruments') that are good predictors of the treatment, but are not independently related to the outcome, may be used to purge the bias. In practice, the validity of the IV approach relies on finding appropriate instruments and these may be hard to find (*see* Jones 2000 and Auld 2006 for further discussion).
- Corrections for selection bias: these range from parametric methods such as the **Heckit** estimator to more recent semiparametric estimators. The use of these techniques in health economics is discussed in Chapter 7.

- Longitudinal data: the availability of **panel data**, giving repeated measurements for a particular individual, provides the opportunity to control for unobservable individual effects which remain constant over time. Panel data models are discussed in Chapter 11.

## Classical linear regression

So far, the discussion has concentrated on the evaluation problem. More generally, most econometric work in health economics focuses on the problem of finding an appropriate model to fit the available data. Classical linear regression analysis assumes that the relationship between an outcome, or dependent variable, *y*, and the explanatory variables or independent variables, *x*, can be summarised by a regression function. The regression function is typically assumed to be a linear function of the *x* variables and of a random error term, ε. This relationship can be written using the following shorthand notation:

$$y = x\beta + \varepsilon. \tag{1}$$

The random error term ε captures all of the variation in y that is not explained by the *x* variables. The classical model assumes that:

- this error term has a mean of zero
- that its variance, $\sigma^2$, is the same across all the observations (this is known as **homoskedasticity**)
- that values of the error term are independent across observations (known as **serial independence**)
- that values of the error term are independent of the values of the *x* variables (known as **exogeneity**).

Often it is assumed that the error term has a **normal distribution**. This implies that, conditional on each observation's $x_i$, each observation of the dependent variable $y_i$ should follow a normal distribution with mean equal to $x_i\beta$.

So far we have not specified how *y* is measured. Often the quantity that is of direct economic interest will be transformed before it is entered into the regression model. For example, data on household healthcare expenditures or on the costs of an episode of treatment have non-negative values only and tend to have highly skewed distributions, with many small values and a long right-hand tail with a few exceptionally expensive cases. Regression analyses of these kinds of skewed data often transform the raw scale, for example by taking logarithms, before running the regression analysis. This reduces the skewness of the distribution and makes the assumption of normality more reasonable. However, the economic interpretation of the results is usually carried out on the original scale, in units of expenditure, and care needs to be taken in retransforming back to this scale. This is particularly true in the presence of **heteroskedasticity**. There is an extensive literature in health economics on this **retransformation problem**, which explores the properties of the logarithmic and other related transformations (*see*, for example, Manning 2006).

In health economics, empirical analysis is complicated by the fact that the theoretical models often involve inherently unobservable (latent) concepts such as health endowments, physician agency and supplier inducement, or quality of life. The widespread use of individual-level survey data means that nonlinear models are common in health economics, as measures of outcomes are often based on qualitative or limited dependent variables. Examples of these nonlinear models include:

- binary responses, such as whether the individual has visited their GP over the previous month (*see* Chapter 3)
- multinomial responses, such as the choice of healthcare provider (*see* Chapters 4 and 5)
- integer counts, such as the number of GP visits (*see* Chapter 9)
- measures of duration, such as the time elapsed between visits (*see* Chapter 10).

Throughout the rest of the book, emphasis is placed on the assumptions underpinning these econometric models and applied empirical examples are provided. The empirical examples are based on a single data set, the Health and Lifestyle Survey of Great Britain (HALS). The next chapter describes how the survey was collected and the kind of information it contains.

Chapter 2

# The Health and Lifestyle Survey

## Survey design

The Health and Lifestyle Survey (HALS) was designed as a representative survey of adults in Great Britain (*see* Cox *et al*. 1987, 1993). The population surveyed was 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 observation, or individual. It was carried out between the autumn of 1984 and the summer of 1985. Information was collected in three stages:

- a one-hour face-to-face interview, which collected information on experience and attitudes towards to 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. 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 wards were selected for each constituency and, finally, 30 addresses per ward. 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, business premises or were derelict (*see* Table 2.1 for details). 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. A total of 9,003 interviews were completed (*see* Table 2.2). This is a response rate of 73.5%. In other words, there was a roughly 1 in 4 chance that an interview was not completed. The missing values are an example of **unit non-response**. For these individuals, no information is available from any of the survey questions. The main reason for non-response is refusal on the part of the interviewee or their family. This accounted for 2,341 cases or 19% of the requests for interview. Further cases were lost because the interviewer was unable to establish contact or for other reasons, such as illness or incapacity on the part of the interviewee.

A question for researchers is whether the 1 in 4 individuals who were not included in the survey are systematically different from those who did respond. If there are systematic differences, this creates a problem of **sample selection bias** and it will not be possible to claim that inferences based on the observed data are representative of the underlying population (*see* Chapter 7). What do we know