THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS
DEPARTMENT OF ECONOMICS & RELATED STUDIES
NHS CENTRE FOR REVIEWS & DISSEMINATION
YORK HEALTH ECONOMICS CONSORTIUM

HEALTH ECONOMETRICS

Andrew M. Jones

HEALTH ECONOMETRICS

Andrew M. Jones

Department of Economics and Related Studies,
University of York,
York, YO1 5DD,
United Kingdom
Tel: +44-1904-433766
Fax: +44-1904-433759
E-mail: amj1@york.ac.uk

Prepared for Chapter 7, *North Holland Handbook of Health Economics*, (eds.) J.P. Newhouse and A.J. Culyer.

This is a preliminary draft and comments are very welcome.

January 1998

CONTENTS

1. Introduction

A decade ago, Newhouse (1987) assessed the balance of trade between imports from the econometrics literature into health economics, and exports from health economics to a wider audience. While it is undoubtedly true that imports of concepts and techniques still dominate the balance, the literature reviewed in this chapter shows that the range and volume of applied econometric work in health economics has increased dramatically over the past ten years. What is more, the prevalence of latent variables, unobservable heterogeneity, and nonlinear models make health economics a particularly rich area for applied econometrics.

The chapter is not a systematic review. Instead, it attempts to provide an overview of the econometric methods that have been applied in health economics, and to use a broad range of examples to illustrate their use. The emphasis of the chapter is on the use of individual level data and microeconometric techniques; reflecting the emphasis on microeconomic analysis in health economics generally. The majority of aggregate analyses have used international data, and the methodological issues surrounding international comparisons of health care are discussed by Jonsson and Gerdtham (1998) in this Handbook.

The structure of the chapter is organised around the nature of the data to be analysed and, in particular, the way in which the dependent variable is defined and measured. This puts the emphasis on the specification of models and appropriate methods of estimation. But the emphasis on estimation should not imply a neglect of checks for model misspecification, and examples of the use of diagnostic tests are given throughout.

2. Identification, heterogeneity, and estimation

2.1 The evaluation problem

The evaluation problem is whether it is possible identify causal effects from empirical data. Mullahy and Manning (1996) provide a concise summary of the problem and, while their discussion focuses on clinical trials and cost-effectiveness analysis, the issues are equally relevant for structural econometric models. An understanding of the implications of the evaluation problem for statistical inference will help to provide a motivation for most of the econometric methods discussed in this chapter.

Consider an "outcome" y_{it}, for individual i at time t; for example an individual's use of primary care services. The problem is to identify the effect of a "treatment" on the outcome; for example whether the individual has health insurance or not. The causal effect of interest is,

$$CE(i,t) = y^{T}_{it} - y^{C}_{it} \qquad (1)$$

where T denotes treatment (insurance) and C denotes control (no insurance). The pure causal effect cannot be identified from empirical data 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 primary care, at time t, both with and without the influence of insurance.

One response to this problem is to concentrate on the average causal effect,

$$ACE(t) = E[y^{T}_{it} - y^{C}_{it}] \quad (2)$$

and attempt to estimate it with sample data. Here it is helpful to think in terms of estimating a general regression function,

$$y = g(x, \mu, \varepsilon) \quad (3)$$

where x is a set of observed covariates, including measures of the treatment, μ represents unobserved covariates, and ε is a random error term reflecting sampling variability. The problem for inference arises if x and μ are correlated and, in particular, if there are unobserved factors that influence whether an individual is selected into the treatment group or how they respond to the treatment. This will lead to biased estimates of the treatment effect.

A randomised experimental design can achieve the desired orthogonality of measured covariates (x) and unobservables (μ); and, in some circumstances, a natural experiment may mimic the features of a controlled experiment (see e.g. Heckman, 1996). However, the vast majority of econometric studies rely on observational data gathered in a non-experimental setting. These data are vulnerable to problems of non-random selection and measurement error which may bias estimates of causal effects.

2.2 Estimation strategies

In the absence of experimental data attention has to focus on alternative estimation strategies. Mullahy and Manning (1996) identify three common approaches:-

i) 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. The debate over whether to treat these unobservables as fixed or random effects, and methods for estimating both linear and nonlinear panel data models are discussed in section 6.

ii) 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. The use of instrumental variables to deal with heterogeneity and simultaneity bias in both linear and nonlinear models is discussed in section 5.

iii) Control function approaches to 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 section 4.3.

Estimation of regression functions like equation (3) typically requires assumptions about the functional form for the deterministic part of the model and about the distribution of the error term. Standard regression analysis assumes that the regression function is linear and that the random error term has a normal distribution. However, in recent years the econometrics literature has seen an explosion of theoretical developments in nonparametric and semiparametric methods which relax functional form and distributional assumptions. These are beginning to be used in applied work in health economics. Section 2.3 introduces kernel-based nonparametric estimators and semiparametric approaches are discussed in sections 4, 6, and 8.

In health economics empirical analysis is complicated further by the fact that the theoretical models often involve inherently unobservable (latent) concepts such as the health endowments, agency and supplier inducement, or quality of life. The problem of latent variables is central to the use of MIMIC models of the demand for health and health status indices (section 5.1.2); but latent variables are also used to motivate nonlinear models for limited and qualitative dependent variables. The widespread use of individual level survey data means that nonlinear models are common in health economics. Examples include binary responses, such as whether the individual has visited their GP over the previous month (section 3.1); multinomial responses, such as the choice of provider (section 3.3); limited dependent variables, such as expenditure on primary care services, which is censored at zero (section 4); integer counts, such as the number of GP visits (section 7); or measures of duration, such as the time elapsed between visits (section 8).

2.3 Nonparametric estimators

Most of the estimators discussed in this chapter rely on assumptions about the functional form of the regression equation and the distribution of the error term. However recent developments in the econometrics literature have focused on semiparametric and nonparametric estimation and many of these are founded on the Rosenblatt-Parzen kernel density estimator. This method uses appropriately weighted local averages to estimate probability density functions of unknown form. Variants on this basic method of density estimation are also used to estimate distribution functions, regression functionals, and response functions [see e.g., Ullah (1988), Duncan and Jones (1992)].

Consider a random vector x with unknown density function f(x). Given a random sample of n observations, the density estimator at a point x is,

$$f_h(x) = (n.\det(H))^{-1} \sum_i K[H^{-1}(x_i - x)] \qquad (4)$$

where K(.) is a multivariate kernel function and det(H) is the determinant of a matrix of bandwidths. Usually the kernel will be a positive real function. In addition, kernel functions are often selected to be symmetric and unimodal density functions. In general, the precise shape of the kernel has little impact on the overall appearance of the density. A central issue in estimation by local smoothing is the choice of bandwidth. Each bandwidth h is a sequence of numbers such that $h \to 0$ and $nh \to \infty$ as the sample size $n \to \infty$. With a fixed sample, the size of h determines the degree of smoothing and is therefore of crucial importance for the appearance, interpretation, and properties of the final estimate. The choice of bandwidth can be a purely subjective choice, it can be based on some rule of thumb, or the choice can be "automated" by data-driven methods such as cross validation.

One feature of the standard kernel estimator is that the size of bandwidth is independent of the point in the sample space at which the estimator is evaluated. This may mean that excessive weight is given to observations in less dense areas of the sample space. The resulting estimates can produce spurious detail, particularly in the tails of estimated densities. Alternative methods are available to overcome this problem; such generalisations distinguish themselves from the basic kernel method by adjusting the bandwidth to account for the density of data in particular regions of the sample space; the less dense the data, the larger the bandwidth. However it should be borne in mind that the greater robustness of these techniques is bought at extra computational cost. Specific methods include the kth nearest neighbour and generalised nearest neighbour, variable kernel, and adaptive kernel methods [see Duncan and Jones (1992)].

Kernel density estimates form the basis for nonparametric regression analysis. In general the regression functional is,

$$E(y|x) = M(x) = \int y f(y|x)dy = \int y(f(y,x)/f(x))\,dy \qquad (5)$$

In nonparametric regression, the regression functional is recovered directly from estimates of the (joint and marginal) density functions. No parametric restrictions are imposed on the form of conditional expectation M(.) or the density function of the implied error term. The Nadaraya-Watson regression estimator takes the form

$$M(x) = \sum y_i W_{hi}(x) \quad , \quad W_{hi}(x) = (n.\det(H))^{-1} K[H^{-1}(x_i - x)]/f_h(x) \qquad (6)$$

The nonparametric regression function is therefore a weighted average, with the individual kernel weights $W_{hi}(x)$ dependent on the estimated kernel density of the regressors.

There appear to have been very few applications of kernel-based nonparametric and semiparametric estimators in health economics. However, as appropriate software becomes more readily available, use of the techniques is likely to increase. Jones (1993) uses data from the 1984 UK Family Expenditure Survey to estimate joint densities and nonparametric regressions for the relationship between household's budget share on tobacco and the logarithm of total non-durable expenditure. Norton (1995) uses kernel estimates to smooth a plot of the fraction of elderly nursing home residents who had "spent-down" at the time of discharge against their time of

discharge. Alderson (1997) uses kernel regressions to investigate the shape of the relationship between health related quality of life (HRQoL) and age, without imposing a functional form on the data. She uses data for the EuroQol, EQ-5D, measure of health status collected as part of the ONS Omnibus survey between January and March 1996. The analysis focuses on inequalities in HRQoL and presents separate regressions for males and females and by occupational social class. Studies by Stern (1996) and Lee et al. (1997), which use kernel based semiparametric estimators of the sample selection model, are discussed in section 4.3.3.

3. Qualitative dependent variables

3.1 Binary responses

Consider a binary dependent variable, y_i, which indicates whether individual i is a "non-participant" or a "participant". In health economics, binary dependent variables have been used to model an extensive range of phenomena; examples include the use of health care services, purchase of health insurance, and starting or quitting smoking.

If the outcome depends on a set of regressors, x, the conditional expectation of y is,

$$E(y_i|x_i) = P(y_i=1|x_i) = F(x_i) \tag{7}$$

In order to estimate (7), F(.) could be specified as a linear function, $x_i\beta$; giving the linear probability model. The linear probability model is easy to estimate, using weighted least squares to allow for the implied heteroscedasticity of the non-normal error term, and may be a reasonable approximation if F(.) is approximately linear over the range of sample observations. However the possibility of predicted probabilities outside the range [0,1] creates a problem of logical inconsistency. A nonlinear specification of F(.) can avoid logical inconsistency.

The most common nonlinear parametric specifications are logit and probit models. These can be given a latent variable interpretation. Let,

$$\begin{aligned} y_i &= 1 \quad \text{iff } y^*_i > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{8}$$

where,

$$y^*_i = x_i\beta + \varepsilon_i$$

and, for a symmetrically distributed error term ε with distribution function F(.),

$$P(y_i=1|\, x_i) = P(y^*_i > 0|\, x_i) = P(\varepsilon_i > -x_i\beta) = F(x_i\beta) \tag{9}$$

Assuming that ε_i has a standard normal distribution gives the probit model, while assuming a standard logistic distribution gives the logit model. These models are

usually estimated by maximum likelihood estimation; the log-likelihood for a sample of independent observations is,

$$\text{LogL} = \sum_i \{(1- y_i)\log(1-F(x_i\beta)) + y_i \log(F(x_i\beta)) \} \qquad (10)$$

Applications of probit, logit, and other models for binary variables are too numerous to list here. One recent example is Buchmeuller and Feldstein's (1997) study of the University of California's decision to impose a cap on its contribution to employees' insurance plans in 1994. This natural experiment allows an analysis of how the resulting change in out-of-pocket premiums affected decisions by UC employees to switch insurance plans. The binary dependent variable indicates whether an employee switched plan, and this is modelled by a latent variable representing the net benefit of switching as a function of the change in premium, plan characteristics, and individual demographic characteristics. Plan switching is estimated using probit models on the full sample of 74,478 employees and for separate types of coverage. Simulations of the change in probability of switching associated with changes in the level of premium show large price effects across all of the models.

3.2 Multinomial and ordered responses

3.2.1 Ordered probits and grouped data regression

The ordered probit model can be used to model a discrete dependent variable that takes ordered multinomial outcomes, e.g. y = 1,2.....,m. A common example is self-assessed health, with categorical outcomes such as excellent, good, fair, poor. The model can be expressed as,

$$y_i = j \quad \text{if} \quad \mu_j < y^*_i \leq \mu_{j+1} \ , \quad j=0,.....m-1 \qquad (11)$$

where,

$$y^*_i = x_i\beta + \varepsilon_i \ , \qquad \varepsilon_i \sim N(0,1) \qquad (12)$$

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,

$$P_{ij} = P(y_i = j) = \Phi(\mu_{j+1} - x_i\beta) - \Phi(\mu_j - x_i\beta) \qquad (13)$$

where $\Phi(.)$ is the standard normal distribution function. With independent observations, the log-likelihood for the ordered probit model takes the form,

$$\text{LogL} = \sum_i \sum_j y_{ij} \log P_{ij} \qquad (14)$$

where y_{ij} is a binary variable that equals 1 if $y_i = j$. This can be maximised to give estimates of β and of the unknown threshold values μ_j. Examples of the use of ordered probit models include Kenkel (1995) who has categorical measures of self-reported

health status and of activity limitation from the Health Promotion/Disease Prevention module of the 1985 U.S. National Health Interview Survey, and Chaloupka and Wechsler (1997) who have a categorical measure of average daily cigarette consumption from the 1993 Harvard College Alcohol Study.

Kerkhofs and Lindeboom (1995) develop an ordered probit model for self-reported health, with state-dependent reporting errors. They are concerned with the potential biases that arise in the use of subjective measures of health when responses are influenced by financial incentives and social pressures. In particular they attempt to isolate the impact of employment status on reporting errors. Their model uses three measures of health. A latent variable, H*, that measures true health; a (categorical) self-reported measure of health, H^s; and an objective measure of health based on professional diagnosis, H^o (in their case the Hopkins symptom checklist). In order to focus on the possibility of state-dependent reporting errors they assume that H^o is a sufficient statistic for the impact of employment status (S) on H*. They assume that observed self-reported health is given by,

$$H^s = j \quad \text{if } \mu_j < H^* \leq \mu_{j+1} \ , \quad j=0,.....m-1 \tag{15}$$

True health is assumed to depend on $f(H^o)$, measured by a set of dummy variables, and demographic characteristics x_1,

$$H^* = f(H^o) + x_1\beta + \varepsilon \ , \qquad \varepsilon \sim N(0,1) \tag{16}$$

and the state-dependent reporting bias is modelled through the threshold values,

$$\mu_j = g_j(S, x_2) \tag{17}$$

These depend on employment status and demographic characteristics x_2. Various specifications of g(.) are used to allow for interactions between employment status and demographics. The typical contribution to the likelihood is,

$$P(H^s = j) = \Phi[g_{j+1}(S,x_2) - f(H^o) - x_1\beta] \ - \ \Phi[g_j(S,x_2) - f(H^o) - x_1\beta] \tag{18}$$

The model is estimated with data on heads of household aged 43-63 from the first wave of the Dutch panel survey (CERRA-I). The sample is split by employment status and ordered probit models are estimated with and without the objective measures of health. This gives evidence of state-dependent reporting bias, identified through interactions between employment status and the demographic variables. Also the results suggest that education influences the way in which people report their health.

Grouped data regression is a variant of the ordered probit model in which the values of the thresholds (μ) 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. Sutton and Godfrey (1995) use grouped data regression to analyse social and economic influences on drinking by young men. Their analysis uses pooled individual data for males aged 18-24 from the British General Household Survey for 1978-1990. As is often the case

with survey measures of alcohol consumption, individuals are assigned to one of seven drinking categories defined by the number of units of alcohol consumed per week, where the range of these intervals is recorded in the survey. They estimate a model in which socio-economic characteristics, along with health-related attitudes and behaviour, are used to predict levels of drinking. A general RESET test for misspecification rejects an OLS specification of the model, but does not reject the grouped data regression. Their results show evidence of a significant interaction between the influence of the price of alcohol and an individual's income.

3.2.2 The multinomial logit

Multinomial models apply to discrete dependent variables that can take (unordered) multinomial outcomes, e.g. y = 1,2.....,m. In health economics this often applies to the choice of insurance plan or health care provider, but could be used to model the choice of treatment regime for an individual patient. It is helpful to define a set of binary variables to indicate which alternative (j=1,....,m) is chosen by each individual (i=1,...,n),

$$\begin{aligned} y_{ij} &= 1 \qquad \text{if } y_i = j \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{19}$$

with associated probabilities

$$P(y_i = j) = P_{ij} \tag{20}$$

With independent observations, the log-likelihood for a multinomial model takes the form,

$$\text{LogL} = \sum_i \sum_j y_{ij} \log P_{ij} \tag{21}$$

The multinomial logit model uses,

$$P_{ij} = \exp(x_i\beta_j) / \sum_k \exp(x_i\beta_k) \tag{22}$$

with a normalisation that $\beta_m = 0$.

Multinomial models are often motivated by McFadden's random utility model. Define individual i's utility from choice j as,

$$U_{ij} = V_{ij}(z_i, x_{ij}) + \varepsilon_{ij} \tag{23}$$

or, in linear form,

$$U_{ij} = z_i\alpha_j + x_{ij}\beta + \varepsilon_{ij} \tag{24}$$

The model assumes that individuals are aware of the unobservable (to the researcher) provider characteristics ε_{ij}, and the individual is assumed to choose the alternative that

gives the maximum utility, so choices are based on net utilities. Typically the ε_{ij} are assumed to be type I extreme value (or Weibull), which has the convenient property that the difference between two EVI variables has a logistic distribution. The multinomial logit can be derived from the random utility model provided that unmeasured attributes ε_{ij}'s are independent. Then,

$$P_{ij} = \exp(z_i\alpha_j + x_{ij}\beta) / \sum_k \exp(z_i\alpha_k + x_{ik}\beta) \qquad (25)$$

giving a tractable closed form solution. Setting $\beta = 0$ gives the multinomial logit or "characteristics of the chooser" model, while setting $\alpha_j = 0$ gives the conditional logit or "characteristics of the choices" model.

The assumption that the ε_{ij}'s are independent implies the independence of irrelevant alternatives (IIA) property,

$$P_{ij} / P_{il} = \exp(z_i\alpha_j + x_{ij}\beta) / \exp(z_i\alpha_l + x_{il}\beta) \qquad (26)$$

So the odds ratio is unaffected by the existence of alternatives other than j and l, (i.e., by changes in the individual's choice set). This implies that if a new alternative is introduced all (absolute) probabilities will be reduced proportionately. Many authors have argued that IIA is too restrictive for many of the applications of multinomial models to health economics. For example, Feldman et al. (1989) argue that, in the case of health insurance plans, the addition of a new plan is more likely to affect the choice of "close substitutes". Much of the recent literature has been concerned with models that relax the IIA assumption such as the nested logit model and the multinomial probit model.

The multinomial logit model can be used in conjunction with two-part models and sample selection models (see Section 4). Haas-Wilson et al. (1988) use data from high option Blue Cross and Blue Shield plans of Federal Employees Benefit Program. The paper makes the case for aggregating health care use by episode of treatment rather than by a fixed period and stresses disaggregation into types of treatment episode; in this case outpatient visits only, outpatient with medication, outpatient with hospitalisation, and hospitalisation only. A two-part specification with a multinomial logit for types of treatment and OLS for levels of expenditure within episodes is used. The results do not find a significant effect of coinsurance rates on types of episode, but there is a significant effect on levels of expenditure.

Haas-Wilson and Savoca (1990) use a Federal Trade Commission survey of contact lens wearers and their suppliers. A multinomial logit is used to estimate effects of both personal and provider characteristics on the choice of providers between opticians, opthamologists, and optometrists. The choice of provider is estimated jointly with quality of care using Lee's method to estimate selectivity corrected regressions for patient outcomes (measured by the "presence of seven potentially pathological eye conditions caused by poorly fitted lenses"). The study finds evidence of selection bias which leads to an overestimate of quality of care provided by opthamologists. The scope for selection bias arises because outcomes depend partly on patients' behaviour, and differences among patients may be correlated with their choice of provider.

3.2.3 The nested multinomial logit

Gertler et al. (1987) investigate the impact of user fees on the demand for medical care in urban Peru, using a 1984 Peruvian household survey. They develop a random utility model in which the demand for medical care is modelled as the decision to seek care and, conditional on that, the decision of which provider to use (public clinic, public hospital, or private doctor). The corresponding econometric specification is the nested multinomial logit model, which relaxes the IIA assumption. The empirical model allows them to predict the revenue consequences and welfare effects of increased user fees, and illustrates the trade-off between efficiency and re-distributive goals. Dor et al. (1987) develop the theoretical model used by Gertler et al. (1987) by including access costs in the budget constraint. They apply the nested multinomial logit model to provider choice using 1985 data from the rural Côte d'Ivoire.

A similar approach is adopted by Feldman et al. (1989) who estimate a model using individual data on the demand for health insurance plans among employees of 17 Minneapolis firms. They argue that the existence of "close substitutes" makes the IIA assumption and, hence, the use of a multinomial logit model unrealistic. The assumption is relaxed by using the nested logit specification which drops the IIA assumption between groups of close substitutes. Freedom of choice of doctor is used to distinguish these health plan nests.

The nested logit model generalises the multinomial/conditional logit as follows. Let $l=1,...,L$ denote "nests" of health plan types. In Feldman et al. there are two nests; IPAs and FFS plans versus HMOs. Within each nest there are $j=1,....,J_l$ plan alternatives. Individual utility is,

$$U_{lj} = w_l \delta + x_{lj} \beta + \varepsilon_{lj} \tag{27}$$

where x_{lj} varies with both the nest and insurance plan, e.g. the premium charged, while w_l varies only with the nest, e.g. freedom to choose a doctor. ε_{lj} is assumed to have a generalised extreme value distribution, which relaxes the assumption that the error terms are independent. Then,

$$P_{lj} = P_l P_{j|l} \tag{28}$$

where,

$$P_{j|l} = \exp(x_{lj}\beta/(1-\sigma))/ \exp(I_l) \tag{29}$$

and,

$$I_l = \log(\Sigma_k \exp(x_{lk}\beta/(1-\sigma)) \tag{30}$$

is the "inclusive value", for nest l. β can be estimated up to the scale factor $1/(1-\sigma)$ by using conditional logit within each nest. Then

$$P_l = \exp(w_l\delta + (1-\sigma)\, I_l) / \sum_l \exp(w_l\delta + (1-\sigma)\, I_l) \tag{31}$$

This shows that ML estimation can be done in two steps. First estimate $\beta/(1-\sigma)$ using conditional logit within each nest, then apply conditional logit across the nests to estimate $(1-\sigma)$, including an estimate of the inclusive value.

Feldman et al. (1989) find that Hausman tests, based on the contrast between conditional and nested logit estimates, suggest that the grouping of IPAs and FFS versus HMOs is satisfactory. But they reject the grouping of IPAs and HMOs. Their results show that health plan choices are sensitive to out-of-pocket payments, and they suggest that estimates of the impact of premiums derived from conditional logit models could be misleading.

3.2.4 The multinomial probit model

The use of a nested logit approach implies that choices can be organised into a meaningful nesting or tree structure. This may not be appropriate for some applications. For example, in their study of the choice of provider between government health facilities, mission health facilities, private clinics and self-treatment in the Meru District of Eastern Kenya, Mwabu et al. (1993) argue that there are no a priori grounds for deciding on the correct decision structure for patients. As a result they adopt the simpler multinomial logit specification, using the IIA assumption.

An alternative to the nested logit model is to use a multinomial probit model. Until recently the computational demands of this model have been prohibitive, but the development of simulation based estimators has opened the way for empirical applications. Bolduc et al. (1996) use data from the rural district of Ouidah in Bénin to model the choice of provider between hospital, community health clinic (CHC), private clinic and self-medication. The empirical focus is on the role of user fees (for the CHCs) and precautionary savings (through tontines) to fund health care. They adopt a random utility specification as in equation (24) and compare multinomial logit (ML), independent multinomial probit (IMP), and multinomial probit (MP) specifications.

The independent probit model assumes that the ε_{ij} are iid normal. Then the probability that individual i chooses j is,

$$P_{ij} = \int_{-\infty}^{\infty} \prod_{k\neq j} \Phi(\, z_i\alpha^*_k + x^*_{ik}\beta + \varepsilon_{ij})\, \phi(\varepsilon_{ij})\, d\varepsilon_{ij} \tag{32}$$

where $\alpha^*_k = \alpha_j - \alpha_k$ and $x^*_k = x_j - x_k$. This specification assumes independence but, unlike the MNL, it does not imply equal cross elasticities (the IIA property).

The multinomial probit model relaxes independence and assumes that the ε_{ij} have a multivariate normal distribution, $N(0, \Omega)$. Then,

$$P_{ij} = \int_{-\infty}^{Ak\neq j} \ldots \int \phi\,(u; \Omega)\, du \tag{33}$$

where $A_k = z\alpha^*_k + x^*_k\beta$. This requires computation of the area under the multivariate normal density $\phi(.)$, such that the utility associated with j is greater than the utility from all the alternatives $k \neq j$.

Bolduc et al. estimate this model using simulated maximum likelihood approach using the GHK simulator. They find that an LR test rejects independence in the probit model. Their estimated time and money price elasticities are sensitive to the empirical specification; those for the multinomial probit are "dramatically different" from those for the multinomial logit and independent multinomial probit. In computing these estimates they use hedonic price and travel time equations based on samples of individuals who use each different provider. This is common practice in the literature (see e.g. Gertler et al. (1987)) but it does raise the issue of potential selection bias.

3.3 Bivariate models

The models discussed in the previous section deal with a single dependent variable that can take multinomial outcomes. The bivariate probit model applies to a pair of binary dependent variables and allows for correlation between the corresponding error terms. It is possible to express the model in terms of latent variables,

$$y^*_{ji} = x_{ji}\beta_j + \varepsilon_j, \quad j=1,2, \quad (\varepsilon_1,\varepsilon_2) \sim N(0,\Omega) \tag{34}$$

where,

$$\begin{aligned} y_{ji} &= 1 \quad \text{iff } y^*_{ji} > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{35}$$

In practice, the health economics literature has made greater use of two variants of the bivariate probit model; the sample selection model and the partial observability probit model. In the model with sample selection y_2 is observed only when $y_1 = 1$. In the partial observability model the researcher observes only $y = y_1.y_2$.

A variant of the partial observability probit assumes that, if $y_1 = 1$, both y_1 and y_2 are observed, while if $y_1 = 0$, then only $y_1.y_2$ is observed. The log-likelihood for this case is,

$$\begin{aligned} \text{LogL} = \sum_{y1=0} \log\Phi(-x_1\beta_1) &+ \sum_{y1=1,y2=0} \log\Phi(x_1\beta_1,-x_2\beta_2,-\rho) \\ &+ \sum_{y1=1,y2=1} \log\Phi(x_1\beta_1,x_2\beta_2,\rho) \end{aligned} \tag{36}$$

In fact, this identical to the bivariate probit with sample selection; and only the interpretation of the model differs. Examples of the application of these models in health economics are van de Ven and van Praag (1981), Kenkel and Terza (1993), van de Ven and van Vliet (1995), and Jones (1993).

The pioneering use of the sample selection model in health economics is van de Ven and van Praag's (1981) study of the demand for deductibles in private health insurance. They use data on 8,000 respondents from a postal survey of 20,000 policy holders of a large non-profit health insurer in the Netherlands, to model choice between a plan with a deductible and one with complete coverage. The dependent variable is derived from a binary response to a question about their preference for a policy with a deductible.

This is modelled as a function of previous use of medical care, self-reported illness days, income, employment and demographics.

Their economic model specifies the expected utility gain from taking a deductible and leads to a basic probit model. However the dataset is prone to selection bias. The survey has a substantial proportion of incomplete responses and these are shown to vary with demographics. van de Ven and van Praag compare a two step estimator with the maximum likelihood estimator of the sample selection model. Incomplete response is predicted by age, gender and family size. Their results show that the two step estimator gives results that are close to ML. They find that health, previous medical consumption and income have significant effects, which implies the potential for adverse selection if individuals can choose between plans with different levels of deductibles.

An example of the partial observability probit model is Kenkel and Terza's (1993) study of the demand for preventive medical care. The motivation for this study is a recognition of the limitations of the neoclassical model of demand for (preventive) medical care, measured by use of diagnostic tests. This stems from the fact that the consumer's (latent) demand is not observed without a visit to doctor, and the actual choice of treatment is influenced by the role of the doctor in mediating patient choice. Together these mean that a physician visit hurdle comes between latent and observable demand for diagnostic tests.

The use of diagnostic tests is modelled as a partial observability probit based on the latent variables,

$$y^*_2 = x_2\beta_2 + w_2\,\alpha_2 + \varepsilon_2 \qquad \text{[diagnostic test index]} \qquad (37)$$

$$y^*_1 = x_1\beta_1 + \varepsilon_1 \qquad \text{[physician visit index]} \qquad (38)$$

Kenkel and Terza's identification strategy relies on the fact that they are modelling sequential decisions. The physician visit is patient initiated, but tests are made after seeing a doctor and are influenced by a set of post-visit influences w_2. Tests for supplier induced demand are based on a sub-set of w_2; those post-visit influences that reflect financial incentives for doctors. Although this is a sequential model, Kenkel and Terza reject a "two-part model", as it rules out positive latest demands for those individuals who do not visit the doctor.

Data from the 1977 National Medical Expenditure Survey is used in separate analyses for men and women and for lab tests and diagnostic tests. The common set of regressors include insurance coverage (private, medicare/caid, none), health (self-assessed and disability days), income, schooling, age, and race. The post-visit variables (w_2) measure outpatient or ER versus office visits, waiting time, and the percentage of the charge paid by private or public insurance. The results show that the correlation between the two error terms is significant for diagnostic tests, but not significant for lab tests. The probability of diagnostic tests increases with private insurance and the fraction of charge paid by private insurance. The results do not support the existence of SID; reflected in the fact that there is no evidence of fewer tests in outpatient/ER compared to office visits, and no effect of waiting time.

4. Limited dependent variables

4.1 Two-part, selectivity, and hurdle models

4.1.1 A taxonomy

Two-part (or multi-part), sample selection, and hurdle models have all been used in the health economics literature to deal with the problem of limited dependent variables. To understand which approach is appropriate for a particular application, it is useful to begin by asking what type of dependent variable is being used. To answer this question it is helpful to introduce some notation. Say that there are two variables of interest: a binary indicator d_i, with associated covariates x_1 and parameters β_1, and a continuous variable y_i, with associated covariates x_2 and parameters β_2, where y_i is coded as $y_i = 0$ if $d_i = 0$.

The first question is whether observations of $y_i=0$ represent an actual choice of zero? If the answer is no, the problem is one of non-observable response and a sample selection model is appropriate (see e.g. Heckman, 1979). For example this might apply to the case where coinsurance rates (y) are only observed for those who purchase insurance (d=1), but non-purchase of insurance does not imply that a potential insuree would face a coinsurance rate of zero. If the answer to the question is yes; then zero observations represent a genuine choice of zero.

In the case of "genuine zeros"; the second question is whether the choice to consume is influenced by the decision of how much to consume? If the answer is no, a sequential decision model is appropriate. If the answer is yes, a joint decision model is appropriate. When considering joint versus sequential decisions it is important to make the distinction between a chronological sequence of events and sequential choice. For example the "gate-keeper" role of GPs may mean that an individual has to visit a GP before they can use inpatient care. This limits their opportunity set, but the individual can consider a range of options; do not visit the GP; visit the GP but do not visit consultant; or visit both. Modelling these decisions as a sequential choice suggests a myopic decision rule, i.e. visit the GP then decide how to respond to advice.

The third question to bear in mind is the object of the analysis. Is the object prediction of $E(y|x)$, inference about β_1 and β_2, inference about $\partial E(y|y>0,x)/\partial x$, etc.? The answer to this question will help to determine the appropriate method to adopt.

Defining the dependent variables in this way suggests a taxonomy to distinguish the three approaches. In the sample selection model, knowledge that y=0 (as opposed to $d_i=0$) is uninformative in estimating determinants of the level of y_i. In the two-part model observations for which $y_i=0$ are uninformative in estimating the determinants of the level of $(y_i| y_i>0)$. In hurdle models, the fact that $y_i=0$ is used in the estimation of β_2 .

It is possible to express the sample selection and hurdle models in terms of latent variables (y*):

$$y^*_{ji} = x_{ji}\beta_j + \varepsilon_j \ , \ j=1,2 \qquad (39)$$

Then the sample selection model is given by;

$$y_{2i} = y^*_{2i} \qquad \text{iff } y^*_{1i} > 0 \qquad (40)$$

$$= \text{unobserved} \quad \text{otherwise } (= 0 \text{ in generalised Tobit})$$

and the hurdle model is given by;

$$y_{2i} = y^*_{2i} \qquad \text{iff } y^*_{2i} > 0 \text{ and } y^*_{1i} > 0 \qquad (41)$$

$$= 0 \qquad \text{otherwise}$$

There is no latent variable representation for the two-part model. Instead it is motivated by a conditional mean independence assumption,

$$E(\varepsilon_{2i} \mid y_{2i} > 0, x_{2i}) = 0 \qquad (42)$$

Notice that no assumption is made about the unconditional mean E(y|x), so the two-part model cannot be used to make inferences about $\partial E(y|x)/\partial x$, only about conditional/selected sample. In general, the two-part specification does not assume normality of $(\varepsilon_1, \varepsilon_2)$ and does not require linearity of E(y|y>0).

4.1.2 Two-part versus selectivity models: the debate

The issue of choosing between the two-part model (2PM) and a generalised Tobit or sample selection specification (SSM) to model the demand for medical care has provoked a vigorous, and often heated, debate in the health economics literature. Advocacy of the two-part model is most associated with the empirical strategy adopted for the RAND Health Insurance Experiment (see e.g. Newhouse et al., 1980, Leibowitz et al., 1985, Manning, Newhouse et al., 1987). Duan et al. (1983) initiated the subsequent debate by making the case for the two-part model. They argue that the censored data approach requires restrictive distributional assumptions and that, as the censored data is unobservable, these assumptions are not testable. They stress "poor numerical and statistical properties" of the SSM, caused by the existence of multiple local optima in its likelihood function. They also argue that the fact that the residual vector is censored in the SSM poses a problem for standard residual based tests.

Hay and Olsen (1984) criticise the 2PM by claiming that it is also subject to untestable assumptions and they question the existence of any distribution of $(\varepsilon_1, \varepsilon_2)$ that gives a complete normal distribution for $(\varepsilon_2|\varepsilon_1 > -x_1\beta_1)$. To support this argument they show that if ε_1 and ε_2 are not independent, the conditional distribution of ε_1 is generally a function of $(x_1\beta_1)$. They respond to the argument that the SSM has poor numerical properties by citing an algorithm for finding a global maximum. Also they argue that, even though the 2PM and SSM are non-nested, they can be compared in terms of

mean squared forecast error (MSFE). Duan et al. (1984) counter this final point by showing that with the RAND data there is no discernable difference between the 2PM and SSM models according to the MSFE criterion. Also they provide an example designed to show that it is possible to find a distribution of $(\varepsilon_1, \varepsilon_2)$ that contradicts Hay and Olsen's claim.

Maddala (1985) sets out to adjudicate the debate. He stresses the need to understand the nature of the underlying decision process in selecting an empirical model and argues that joint decisions may be more appropriate than the sequential approach implied by the 2PM. He cites van de Ven and van Praag (1981) and argues that decision to use health care will be linked to perceived severity of illness (and hence likely expenditure).

In response to Duan et al. (1984) he points out that semiparametric estimators were available for the SSM and that the normality assumption is testable. Also he considers their "counter-example". Duan et al. (1984) aim to show that there is a joint distribution of $(\varepsilon_1, \varepsilon_2)$ that allows correlation between the two error terms but, for d=1, gives,

$$\log(y) = x_2\beta_2 + \varepsilon_2, \varepsilon_2 \sim IN(0, \sigma^2) \qquad (43)$$

They assume that ε_1 is continuous for the whole population, and that ε_2 has a mass point at $\varepsilon_2=-\infty$ and is continuous over the real line for d=1. They argue that it is possible to construct a joint distribution from these marginals such that ε_1 and ε_2 are correlated. Maddala argues that this is "purely semantic" as the correlation is not estimable. Also, their model is actually specifying conditional distributions for the separate sub-populations, $\varepsilon_1 > x_1\beta_1$ and $\varepsilon_1 \leq x_1\beta_1$.

Maddala makes the distinction between sample selection models; in which the criterion function is written in reduced form, and correlation between ε_1 and ε_2 is the only connection between the two equations; and self selection models in which the criterion is written in structural form. He argues that adopting a structural approach "will help in organising one's thinking properly on why one expects any selectivity bias in the problem". He goes on to argue that "even when decisions are sequential, if there are some common omitted variables the two decisions will be correlated. In this case, it is advisable not to formulate the model in a way that the correlation cannot ever be estimated". Zimmerman Murphy (1987) lists common omitted variables in context of medical care demand; these include insurance status, time costs, marginal valuation of health, time preference, and risk aversion.

Duan et al. (1985) take up Maddala's challenge. They stress that the focus of their own work is on estimating mean medical expenditure and that, in that context, the debate over statistical methods has no relevance for the policy implications of their results. They find that multi-part, ANOVA, and sample selection models all give similar results, and that the debate is "much ado about nothing". Also they argue that "in the specific case of health insurance one does not need an estimate of ρ to estimate mean expenditure" and that many econometrics models are formulated so that "nuisance parameters" are eliminated, these include the Cox partial likelihood, the

within-groups estimator for panel data, and zero restrictions in structural models. Maddala (1985) rounds off the exchange by recognising that the RAND data is special because participants were randomised across insurance plans. But he cautions against use of the 2PM in other contexts.

4.1.3 Monte Carlo evidence

In an attempt to settle the debate over the relative merits of 2PM and SSM specifications Manning, Duan, and Rogers (1987) use Monte Carlo simulations to compare the LIML (Heckit) and FIML sample selection estimators with a "naive two-part model" (the true specification omitting the correlation coefficient) and a "data-analytic (testimator) variant" (which adds powers of x, according to a test criterion).

As in the earlier work of the RAND researchers, they stress that "we are not interested in the coefficients per se", only in predictions of E(y) using

$$E(y) = P(y>0)E(y|\, y>0) \tag{44}$$

They use the SSM as their theoretical benchmark, but find that 2PM outperforms it on statistical grounds. This leads them to conclude; "based on our experience here and elsewhere, we believe that the data-analytic version of the two part model will be robust - as long as analysts are concerned about the response surface rather than particular coefficients".

A comprehensive re-assessment of the Monte Carlo evidence in Manning, Duan, and Rogers (1987) is provided by Leung and Yu (1996). Leung and Yu argue that their Monte Carlo design creates collinearity problems that bias the results against the SSM and in favour of 2PM. The design problem they identify is that Manning, Duan, and Rogers use a model with no exclusion restrictions ($x_1 \equiv x_2$) and simulate $x \sim u(0,3)$. Leung and Yu argue that this leads to insufficient range of variation in the inverse Mill's ratio. Leung and Yu use $x \sim u(0,10)$ and find that "collinearity problems vanish and the sample selection model performs much better than the two-part model". Of course this raises the empirical question of how much variation will be observed with real data.

To understand the collinearity problem consider the Heckit/LIML estimator of the SSM, which is based on,

$$y = x_2\beta_2 + \lambda(x_1\beta_1) + \varepsilon_2 \tag{45}$$

With $x_1 \equiv x_2$, identification (of β_2) relies on the nonlinearity of the inverse Mill's ratio $\lambda(.)$. A plot of $\lambda(.)$ shows that the function is approximately linear for much of its range. This implies that the range of $x_1\beta_1$, and hence of x_1, is important and that the degree of censoring is important, as it reduces the range of observed values. Leung and Yu argue that the claim that Heckit will perform poorly when there is a high degree of correlation between $x_1\beta_1$ and x_2 is potentially misleading. In their Monte Carlo design, Heckit performs well when x_2 and $x_1\beta_1$ are perfectly correlated, as long as the

proportion of censored observations is sufficiently small and/or the range of x_1 is sufficiently large (i.e. when the nonlinearity of $\lambda(.)$ comes into play).

Leung and Yu (1996) conclude that the performance of models depends on the empirical context. Collinearity problems can arise if there are few exclusion restrictions, a high degree of censoring, low variability among the regressors (x_1), or a large error variance in the choice equation (i.e. weak instruments). They suggest that applied researchers should always check for collinearity. After looking at a range of measures of collinearity, they favour the condition number. They argue that their Monte Carlo evidence shows that, in the absence of collinearity problems, the t-test on the inverse Mill's ratio can be used to distinguish between the 2PM and SSM. Overall they conclude that "....the merits of the two-part model have been grossly exaggerated in the literature".... "hence the extreme and negative remarks against the sample selection model made by Duan et al. ..... are unwarranted and misleading".

4.1.4 Empirical evidence

Zimmerman Murphy (1987) estimates sample selection models for physician office visits, hospital outpatient visits, and hospital inpatient days using the 1970 U.S. National Health Survey. She uses the Heckit estimator and finds significant negative coefficients for the inverse Mill's ratio. The results show evidence of the collinearity problem, with the estimates of the selectivity correction becoming less significant the greater the correlation between the inverse Mill's ratio and the other regressors . Hunt-McCool et al. (1994) use a sample of adult from the U.S. National Medical Care Expenditure Survey. Their dependent variables are the quantity of service (office visits, hospital inpatient care) and out-of-pocket expenditure shares. Heckit estimates show positive and significant coefficients on the inverse Mill's ratio.

4.2 Two-part models: developments and applications

This section draws heavily on John Mullahy's (1997) paper "Much ado about two: reconsidering the two-part model in health econometrics". Mullahy focuses on the 2PM applied to "genuine zeros" rather than missing observations. He argues that, due to nonlinearities and retransformations, the estimated parameters from the 2PM are not sufficient for inference about important policy parameters that involve the level oy y, such as $E(y|x)$, $\partial E(y|x)/\partial x$, and $\partial \log E(y|x)/\partial \log x$.

The usual specification of the 2PM separates a probit or logit for $\pi(x)=P(y>0|x)$, and least squares estimates on the logarithm of y,

$$\begin{aligned} \log(y) &= \log(\mu(x)) + \varepsilon_2, \quad y > 0 \\ &= x\beta_2 + \varepsilon_2 \end{aligned} \tag{46}$$

The problem for inference stems from two issues; the conditioning on y>0, and the need to re-transform from log(y) to y-space. The identifying assumption for β_2 is the

orthogonality condition $E(\varepsilon_2|y>0,x)=0$. Under this assumption the 2PM will give consistent estimates of β_2, but the condition does not identify other parameters such as E(y|x). In general notation, the 2PM implies,

$$\begin{aligned} E(y|x) &= P(y>0|x).\, E(y|y>0,x) \\ &= \pi(x).\mu(x).\, E(\exp(\varepsilon_2)|\, y>0,x) \\ &= \pi(x).\mu(x).\rho(x) \end{aligned} \tag{47}$$

with parametric representations,

$$= \pi(x;\beta_1).\mu(x;\beta_2).\rho(x;\gamma) \tag{48}$$

The presence of $\rho(x;\gamma)$ in this expression means that the identification of (β_1,β_2), by the 2PM, is not sufficient to identify E(y|y>0,x) or E(y|x). Two solutions to this identification problem are:

1) Assume log-normality of (y|y>0,x) with constant variance σ^2, which implies,

$$E(y|y>0,x) = \exp(x\beta_2 + \tfrac{1}{2}\sigma^2) \tag{49}$$

2) Instead of assuming a distribution for ε_2, Duan (1983) proposes a nonparametric smearing estimator,

$$S = \sum^{n+}_{i=1} [\exp(\varepsilon_{2i})]/\, n_+ \tag{50}$$

the mean of the estimate of $\exp(\varepsilon_{2i})$ over the positive observations (n_+). Duan shows that this is a consistent nonparametric estimator of $E(\exp(\varepsilon_2))$.

The problem with the smearing estimator is that consistent estimation of β_2 in the 2PM only requires the orthogonality condition $E(\varepsilon_2|y>0,x) = 0$. In other words ε_2 could be heteroscedastic, in which case the consistency of the smearing estimation breaks down. Mullahy (1997c) speculates about testing for this problem by running a regression of $\exp(\varepsilon_2)$ on, say, $\exp(x\gamma)$. If $\gamma=0$ for non-constant elements of x, then Duan's estimator should be adequate. The actual approach adopted by the RAND researchers was to split the sample by discrete x variables and apply separate smearing estimates.

Given the problems of identifying E(y|y>0,x) or E(y|x) in the standard 2PM, Mullahy (1997c) considers two alternative estimators. First, given that E(y|y>0,x) must be positive, he suggests using an exponential conditional mean specification; $E(y|y>0,x)=\exp(x\beta_2)$. Combining this with a logistic specification for P(y>0|x), the model gives,

$$\begin{aligned} E(y|x) &= P(y>0|x).\, E(y|y>0,x) \\ &= [\exp(x\beta_1)/(1+\exp(x\beta_1))].\exp(x\beta_2) \\ &= \exp(x(\beta_1+\beta_2))/(1+\exp(x\beta_1)) \end{aligned} \tag{51}$$

The model can be estimated by a two-step estimator (2PM-M-2); using logit (or probit) for β_1, and nonlinear least squares (NLLS) for the positive observations. Alternatively it can be estimated in one step (2PM-M-1), using the full sample to estimate (51) by NLLS.

Then Mullahy considers the "more primitive assumption", $E(y|x) > 0$. This can be justified by the fact that, for non-negative y, finding $E(y|x)=0$ means the problem is uninformative (as it implies that y always equals zero). So he suggests using the exponential conditional mean (ECM) model, $E(y|x)=\exp(x\beta)$, and estimating by NLLS. The advantages of this simple specification are that it is straightforward to use instrumental variables to deal with problems of unobservable heterogeneity in the model, and that the elasticities, $\partial E(\log y)/\partial \log x$, are simple to compute and interpret. The price of using the simpler specification is that it does not allow separate inferences about $P(y>0|x)$ and $E(y|y>0,x)$. Mullahy notes that the 2PM-M model reduces to the ECM model when $\beta_1=0$, and he proposes a conditional moment test to assess whether a one-part or a two-part specification applies. He also proposes a Wald test based on the contrast between the 2PM and 2PM-M estimates of β. This can be interpreted as a test of whether $\rho(x)$ is constant.

It is worth noting that the use of exponential conditional mean specifications provides a direct link with the count data regressions discussed in Section 7 of this chapter. The ECM model corresponds to a Poisson regression model, while the 2PM-M-2 corresponds to the the zero altered Poisson model. These specifications are discussed in greater detail below.

4.3 Selectivity models: developments and applications

4.3.1 Manski bounds

In a recent review, Manski (1993) argues that "the selection problem is, first and foremost, a failure of identification. It is only secondarily a difficulty in sample inference." To illustrate, consider a population characterised by (y,d,x), where d and x are observed but y is only observed if d=1. Interest centres on the unconditional probability,

$$P(y|x) = P(y|x,d=1)P(d=1|x) + P(y|d=0,x)P(d=0|x) \tag{52}$$

The selection problem stems from the fact that the term $P(y|d=0,x)$ cannot be identified from the available data. All that is know is,

$$P(y|x) \in [\, P(y|x,d=1)P(d=1|x) + \gamma P(d=0|x)\,,\ \gamma \in \Gamma\,] \tag{53}$$

where Γ is the space of all probability measures on y. A common response to this problem in the statistical literature is the assumption of independence or ignorable non-response,

$$P(y|x) = P(y|d=0,x) = P(y|d=1,x) \tag{54}$$

But, as Manski points out, "in the absence of prior information this hypothesis is not rejectable"; to see this set $\gamma = P(y|d=1,x)$. So, in the absence of prior information, the

"selection problem is fatal for inference on the mean regression of y on x". Restrictions on P(y|x), P(y|x, d=0), and P(d|x,y) may have identifying power, but restrictions on P(y|x,d=1) and P(d|x) are "superfluous" as they are already identified by censored sampling process.

The selection problem may be fatal for inferences concerning E(y|x) without indentifying restrictions, but Manski shows that it is possible to put bounds on other features of the distribution. Let g(.) be a function mapping Y into a bounded interval $[K_0,K_1]$, for example the CDF for y. Then,

$$E[g(y)|x] = E[g(y)|d=1,x]P(d=1|x) + E[g(y)|d=0,x]P(d=0|x) \quad (55)$$

In which case E[g(y)|d=0,x] is not identified by the data, but it is bounded. This implies that E[g(y)|x] can be bounded to a bandwidth which is proportional to $(K_1 - K_0)P(d=0|x)$. "Therefore meaningful to say that degree of underidentification of $E[g(y)|x=x_0)$ is proportional to censoring probability at x_0." This leads Manski to discuss the (nonparametric) estimation of the bounds and to develop estimators for quantile regressions. Quantile regression has been applied by Manning et al. (1991) to analyse whether heavy drinkers are more or less responsive to the price of alcohol than other drinkers. They find evidence that the price effect does vary by level of consumption.

4.3.2 The propensity score

The propensity score approach to dealing with the selection problem has been developed in the context of the identification of treatment effects when there is a problem of self-selection in the assignment of patients to treatments. Rosenbaum and Rubin (1983) show that conditioning on the propensity score, which measures the probability of treatment given a set of covariates, can control for confounding by these covariates in estimates of treatment effects.

Angrist (1995) provides weak sufficient conditions for conditioning on the propensity score in a general selection problem involving instrumental variables. The main identifying assumption is that the instruments satisfy a simple monotonicity condition, as in Imbens and Angrist (1994). The result implies that, with P(d=1|x) fixed, selection bias does not affect IV estimates of slope parameters. This result lies behind Ahn and Powell's (1993) approach to the selection problem, which uses differencing of observations for which non-parametric estimates of P(d=1|x) are "close". To illustrate it is worth recapping a general version of the sample selection model. Assume that the following is observed,

$$y = [x_2\beta_2 + \varepsilon_2]\ \mathbf{1}[\varepsilon_1 > -\psi(x_1)] \quad (56)$$

where $\mathbf{1}[.]$ is an indicator function, $\psi(x_1)$ is the selection index and d is the observed binary variable, such that $d=\mathbf{1}[\varepsilon_1 > -\psi(x_1)]$. For the selected sample,

$$E[y|x,d=1) = x_2\beta_2 + E[\varepsilon_2|x, \varepsilon_1 > -\psi(x_1)] \quad (57)$$

If the distribution of $(\varepsilon_1, \varepsilon_2)$ is independent of x_1 and x_2, the conditional expectation of ε_2 depends only on $\psi(x_1)$. This gives a "tiered index structure",

$$E[y|x] = G_1[x_2\beta_2, \psi(x_1)] \quad (58)$$
$$E[d|x] = G_2[\psi(x_1)] \quad (59)$$

The propensity score is defined as follows,

$$P(x_1) = P(d=1|x_1) = P[\varepsilon_1 > -\psi(x_1)] \quad (60)$$

When the function is independent of x, it is invertible and it is possible to write $\psi(x_1)=\eta(P(x_1))$. Then,

$$E[y|x,d=1) = x_2\beta_2 + \tau[P(x_1)] \quad (61)$$

Using a "differenced form" this leads to ,

$$y - E[y|P(x_1)] = [x_2 - E(x_2|\, P(x_1))]\beta_2 + e \quad (62)$$

where $E[e|x,d=1] = 0$. This forms the basis for the estimators discussed below.

4.3.3 Semiparametric estimators

Ahn and Powell (1993) propose an estimator for the general model where the selection term depends on the propensity score. Consider any pair of observations where $P_i \approx P_j$. Then, provided the selection function $\tau(.)$ is continuous,

$$y_i - y_j \approx [x_{2i} - x_{2j}]\, \beta_2 + \varepsilon_{ij} \quad (63)$$

This leads Ahn and Powell to suggest a weighted IV estimator for β_2, using kernel estimates of (P_i-P_j) as weights,

$$w_{ij} = (1/h)\, K[\, (P_i - P_j)/h]d_i d_j \quad (64)$$

The attraction of this approach is that it gives a two-stage estimation procedure with closed form solutions at both stages. The first step is to construct a standard kernel regression for the propensity score P based on the observed d's. Estimates of β are given by the weighted IV estimator. Ahn and Powell show that, under appropriate assumptions, the estimator is $\sqrt{n}$ consistent and asymptotically normal and they provide an estimator for the associated covariance matrix.

The Ahn and Powell approach is particularly flexible because it is based on $\tau[P(x_1)]$. Many other semiparametric approaches have concentrated on the linear index version of the selectivity model,

$$E[y|x,d=1) = x_2\beta_2 + \lambda(x_1\beta_1) \quad (65)$$

Stern (1996) provides an example of the semiparametric approach in a study that aims to identify the influence of health, in this case disability, on labour market participation. The paper uses a Heckman style model, using labour market participation to identify the reservation wage (supply) and a selectivity corrected wage equation to identify the offered wage (demand). This proves to be sensitive to distributional assumptions and exclusion restrictions.

Stern's data are a sample of 2,674 individuals from the 1981 U.S. Panel Study on Income Dynamics. Disability is measured by a limit on the amount or kind of work the person can do. Initial estimates are derived from reduced form probits and selectivity corrected reduced form wage equations. He finds that disability is insignificant when controlling for selection but very significant without control (even though the selection term is not significant); a result which seems to highlight the collinearity problems associated with the sample selection model. Structural participation equations, in the form of multiple index binary choice models, were very sensitive to the choice of exclusion restrictions, so Stern turns to semiparametric estimation.

He uses Ichimura and Lee's (1991) estimator for the model,

$$y = z_o + \psi(z_1, z_2) + \varepsilon \tag{66}$$

where $z_j = x_j\beta_j$. This includes two special cases that are relevant here: first the structural participation model, where $\beta_0 = 0$, z_1 is the demand index, and z_2 is the supply index; and second the Heckman wage equation, where $z_2=0$. Ichimura and Lee's approach uses a semiparametric least squares (SLS) estimator and minimises the criterion,

$$(1/n) \sum [(y - z_0) - E(y-z_0|z_1,z_2)]^2 \tag{67}$$

where the conditional expectation is given by the nonparametric regression function,

$$\begin{aligned} E(y-z_0|z_1,z_2)] = (1/n-1) &[\textstyle\sum_{j\neq i}(y_j - z_{0j})K[(z_{1i} - z_{1j})/h_1, (z_{2i} - z_{2j})/h_2] \\ &/(1/n-1) \textstyle\sum_{j\neq i} K[(z_{1i} - z_{1j})/h_1, (z_{2i} - z_{2j})/h_2] \end{aligned} \tag{68}$$

and where K[.,.] is a kernel function and the h's are bandwidths. The IL estimator is known to be badly behaved in small samples. In Stern's application this shows up in the irregular shape of the estimated supply function. To deal with this he imposes a monotonicity assumption; ψ_1 , $\psi_2 \geq 0$.

For the multiple index model he reports the correlations for the regressors that are common to both equations. He finds a low degree of correlation and concludes that the "hypothesis that demand and supply are not identified can be rejected" (p.61). The results suggest that the supply effects of disability are much greater than the demand effects. "Thus effort to improve the handicap accessibility of public transportation or home care programmes for disabled workers (if effective at reducing the supply index) are likely to be more successful than efforts to reduce discrimination among employers or to provide wage subsidies to employers" (p.68).

Similar semiparametric methods are used by Lee et al. (1997). Like Stern (1996), they adopt a linear index specification and use semiparametric estimators to avoid imposing any assumptions on the distributions of the error terms in their model. Their analysis is concerned with estimating a structural model for anthropometric measures of child health in low income countries. They argue that reduced form estimates of the impact of health interventions, such as improved sanitation, on child health may be prone to selection bias if they are estimated with the sample of surviving children. If the health intervention improves the chances of survival it will lower the average health of the surviving population, as weaker individuals are more likely to survive, and lead to a biased estimate of the effectiveness of the intervention.

They specify a system of structural equations. These consist of a survival equation, based on a binary dependent variable, which includes the influence of water supply and sanitation on child survival; reduced form input demands, measuring calorie intake; and the child health production function, measured by the child's weight. The survival equation is specified as a linear index model with an unknown error distribution, and is estimated by a semiparametric maximum likelihood (SML) procedure. The reduced form input demands, for the surviving children, are estimated as sample selection models by semiparametric least squares (SLS), conditioning on the SML estimates of the survival index. The child health (weight) production function is estimated using the same approach, but the endogenous health inputs are replaced by fitted values from SLS estimates of the reduced forms, giving two-stage semiparametric least squares estimates (TSLS). The form of the kernel functions and the bandwidths used in the estimation are selected so that the semiparametric estimates are $\sqrt{n}$-consistent and asymptotically normal. Hausman type tests are used to compare the SML estimates of the survival equation with standard probit estimates, to test for the exogeneity of the health inputs, and to test whether there is a problem of sample selection bias.

The models are estimated on two datasets; the 1981-82 Nutrition Survey of Rural Bangladesh and the 1984-85 IFPRI Bukidnon, Philippines Survey. The data are split into sub-samples for children aged 1-6 and 7-14. Tests for normality in the survival equation fail to reject the standard probit model in both of the sub-samples for the Philippines, and for ages 1-6 in Bangladesh. For children aged 7-14 in Bangladesh the estimated effects of maternal schooling and water supply are substantially different, but the estimates for other variables are similar for SML and the probit. For the health production functions they compare a standard simultaneous equations estimator, a simultaneous equations selection model based on joint normality, and the semiparametric estimator. The results do not appear to be sensitive to either the selectivity correction or the normality assumption. Despite this, the authors note that previous reduced form studies may have understated the impact of health interventions, because of the unobservable heterogeneity bias associated with a reduced allocation of resources to child health in households with better facilities.

4.3.4 Identification by covariance restrictions

Pitt (1996) develops similar theoretical ideas to Lee et al. (1997), but adopts a different approach to dealing with the selection problem. He argues that fertility selection bias (when parents are influenced by health prospects for potential births) may affect

estimates of the determinants of child health and mortality, and that mortality selection bias may influence the analysis of the determinants of child health. This creates an identification problem for standard parametric approaches to the selectivity problem. Pitt argues that in a reduced form specification of child health (mortality) conditional on fertility it is difficult to justify exclusion restrictions, that is to find regressors that influence fertility choices but do not influence child health. In this case identification would have to rely on nonlinearity of the selection correction. For binary measure of child health (e.g. mortality data) this leads to a bivariate probit with partial observability, and Pitt cites the empirical problems of identifying this model with his data, and in other studies.

Pitt suggests an alternative approach based on identification by covariance restrictions. This provides a strategy for identification "so long as fertility and health outcomes are observed for more than one time period in the life of each woman in the sample". In other words this approach relies on longitudinal data to control for individual effects. Pitt models observed births (F) and deaths (D) in terms of latent variables,

$$F^*_{it} = x_{fit}\beta_f + \mu_{fi} + \varepsilon_{fit} \quad , \qquad F=1 \quad \text{if } F^*>0 \tag{69}$$

$$D^*_{it} = x_{hit}\beta_h + \mu_{hi} + \varepsilon_{hit} \quad , \qquad D=1 \quad \text{if } D^*>0 \tag{70}$$

Identification relies on there being individual effects that influence fertility (μ_{fi}) which are correlated with the individual effects that influence child health (μ_{hi}). Pitt uses longitudinal data on births to identify these correlated effects. The model adopted is a random effects bivariate probit, which implies that correlation only works through a time invariant effect, i.e. there are no dynamic effects associated with the timing of births.

The model is applied to data from 14 Sub-Saharan Demographic and Health Surveys (DHS). The measure of mortality is deaths before age two. For each country Pitt compares standard probits on the sample of all births, a random effects probit, and a "selection corrected probit", i.e. a random effects bivariate probit with partial observability. There is evidence of correlated effects in all cases. But the random effects only account for a small portion of overall error variance, and there is no marked effect on the derivative of the conditional probability of infant death with respect to parental education.

Pitt also derives trivariate models for continuous anthropometric measures of child health: weight and height. To observe these measures the child must be born and survive and estimation must allow for both sources of selection bias. Estimates from the Zambian DHS show little evidence of selection bias for log(weight) and only limited evidence for log(height).

4.4 Hurdle models: developments and applications

In health survey data, measures of continuous dependent variables such as alcohol and tobacco consumption, or measures of medical care expenditure invariably contain a high proportion of zero observations and appropriate limited dependent variable techniques are required. The special feature of the double hurdle approach is that, unlike the standard Tobit model, the determinants of participation (e.g., whether to start or quit smoking) and the determinants of consumption (e.g., how many cigarettes to smoke) are allowed to differ.

However, a limitation of the standard double hurdle specification is that it is based on the assumption of bivariate normality for the error distribution. Empirical results will be sensitive to misspecification, and ML estimates will be inconsistent if the normality assumption is violated. This may be particularly relevant if the model is applied to a dependent variable that has a highly skewed distribution, as is often the case with survey data on cigarette and alcohol consumption, and for medical care expenditure.

A flexible generalisation of the double hurdle model is used by Yen and Jones (1996). The Box-Cox double hurdle model provides a common framework that nests standard versions of the double hurdle model and also includes the generalised Tobit model and 'two-part' dependent variable, as special cases. This allows explicit comparisons of a wide range of limited dependent variable specifications that have been used in the health economics literature. The model for the observed dependent variable (y_i) can be written in terms of two latent variables (y^*_1, y^*_2), where,

$$y^*_{ji} = x_{ji}\beta_j + \varepsilon_j, \quad j=1,2, \tag{71}$$

$$(\varepsilon_1,\varepsilon_2) \sim N(0, \Omega) \text{ and } \Omega = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma^2 \end{bmatrix}$$

and,

$$\begin{aligned} y^*_{2i} &= (y^{\lambda} - 1)/\lambda \quad \text{for } \lambda>0 \\ &\qquad\qquad \text{iff } y^*_{1i} > 0 \text{ and } y^*_{2i} > -1/\lambda \\ &= \log(y_i) \quad \text{for } \lambda=0 \\ &= 0 \qquad\qquad \text{otherwise} \end{aligned} \tag{72}$$

In other words, the conditional distribution of the latent variables is assumed to be bivariate normal. This specification allows participation to depend on both sets of regressors x_{1i} and x_{2i} and permits stochastic dependence between the two error terms. In addition, the use of the Box-Cox transformation relaxes the normality assumption on the conditional distribution of y_i. Yen and Jones (1996) show that the log-likelihood function for a sample of independent observations is,

$$\begin{aligned} \text{LogL} &= \textstyle\sum_{y=0} \log[1 - \Phi(x_1\beta_1, (x_2\beta_2 + 1/\lambda)/\sigma, \rho)] \\ &+ \textstyle\sum_{y>0} \log\Phi\,[(x_1\beta_1 + (\rho/\sigma)\{(y^{\lambda} - 1)/\lambda - x_2\beta_2\})/\sqrt{(1-\rho^2)}] \\ &+ \textstyle\sum_{y>0} (\lambda - 1)\log(y_i) + \sum_{y>0} \log[(1/\sigma)\phi\,(\{(y^{\lambda} - 1)/\lambda - x_2\beta_2\}/\sigma)] \end{aligned} \tag{73}$$

where Φ denotes a univariate or bivariate standard normal CDF, ϕ denotes the univariate standard normal PDF, and $\rho=\sigma_{12}/\sigma$. The general model can be restricted to give various special cases:

i) $\sigma_{12} = 0$ gives the Box-Cox double hurdle with independent errors.

ii) $\lambda = 1$ gives the standard double hurdle with dependence. This model is applied to UK data on household tobacco expenditure from the 1984 Family Expenditure Survey (FES) in Jones (1992), and to Spanish Family Expenditure Survey data for 1980-81 in Garcia and Labeaga (1996). The special case in which the error terms are assumed to be independent is applied to FES data on household tobacco expenditure in Atkinson et al. (1984), UK data on individual cigarette consumption from the 1980 General Household Survey (GHS) in Jones (1989), and to US data on wine consumption in Blaylock and Blisard (1993).

iii) With $\lambda = 0$ the likelihood function corresponds to the generalised Tobit model with $\log(y_i)$ as dependent variable in the regression part of the model. Setting $\sigma_{12}=0$ gives the special case of the two-part model in which normality is assumed and the equations are linear. Studies of smoking based on the two-part model include Lewit et al. (1981), Wasserman et al.(1991), and Blaylock and Blisard (1992).

Yen and Jones (1996) apply the Box-Cox double hurdle model to data on the number of cigarettes smoked in a sample of current and ex-smokers from the British Health and Lifestyle Survey. The estimated Box-Cox parameter (λ) equals 0.562 which is significantly different from both zero and one at the 0.01 level. Thus, both the standard double hurdle model and generalised Tobit model are rejected.

5. Unobservable heterogeneity and simultaneous equations

5.1 Linear models

5.1.1 Instrumental variables

Problems of unobservable heterogeneity bias and simultaneity have received particular attention in the context of empirical studies of health production. A pioneering paper is Auster et al.'s (1969) analysis of cross sectional data on death rates across the United States in 1960. They specify a Cobb-Douglas model for mortality rates, as a function of medical care and environmental variables. This is estimated by two-stage least squares (2SLS) to allow for the possible endogeneity of medical care; recognising that aggregate mortality rates may influence the level of spending on medical care at the State level.

Rosenzweig and Schultz (1983) highlight the problem of unobservable heterogeneity bias in a study of child health production and the demand for child health inputs. They consider estimation of a structural health production function,

$$y = f(x,z,\mu) \tag{74}$$

where y is a measure of child health, x are goods that affect health such as nutrition, z is medical care, and μ is an unobservable (to the researcher) variable reflecting the child's genetic and enviromental endowment. If the child's parents are aware of μ, it may influence the reduced form demands for health inputs; for example, a mother who has a history of complications during previous pregnancies may be more likely to seek early prenatal care. Then the marginal effect of medical care on health is,

$$\partial y/\partial z = f_z + f_\mu \partial \mu/\partial z \tag{75}$$

So, estimates that fail to control for μ will give biased estimates of the effect of medical care on health (f_z). Rosenzweig and Schultz's (1983) proposed solution is to find instruments that predict the use of medical care but do not have an independent effect on health outcomes, and to estimate the model by 2SLS. Data on live births from the U.S. National Natality Followback Surveys for 1967-69 are used and separate models are estimated for birth weight, the length of the gestation period, and the fetal growth rate. Estimates of the impact of the delay before the mother sought medical care change significantly when 2SLS is used rather than OLS.

A similar static health production framework is adopted by Mullahy and Portney (1990) to estimate the impact of smoking and atmospheric pollution on respiratory health. They use individual data from the 1979 U.S. National Health Interview Survey, and models are estimated for a binary dependent variable indicating whether the individual experienced days when their activities were limited by respiratory illness, and for the actual number of restricted activity days. Both models are estimated by OLS and by the generalised method of moments (GMM); where the latter uses the price of cigarettes and additional demographic variables to instrument the measure of cigarette smoking and the estimation uses 2SLS with a Huber-White correction for heteroscedasticity. In order to assess the sensitivity of the results to the use of instrumental variables, the models are estimated on different sub-samples and with different instrument sets. The results appear to be robust and show that allowing for unobservable heterogeneity bias increases the estimated impact of smoking relative to the impact of atmospheric pollution.

Mullahy and Sindelar (1996) extend the use of GMM estimation of the linear probability model to a two equation system in which a measure of problem drinking is treated as an endogenous regressor in equations for employment and unemployment (with non-participation in the workforce as the omitted employment status). The study is careful to acknowledge the possibility of IV bias, which arises if the instruments are poor predictors of the endogenous regressor, and it reports F-statistics for the significance of the instruments in the reduced form regressions. Data from the 1988 Alcohol Supplement of the NHIS are used and the estimates show that problem drinking has a negative effect on employment.

5.1.2 The MIMIC model

In models of the demand for health and of health status indexes, problems of endogeneity are compounded by the fact that the central concept, "health", is inherently unobservable and has to be proxied by indicator variables. Multiple causes-multiple indicators, or MIMIC, models have been widely used to deal with the problem of latent variables. MIMIC models are estimated as LISREL (linear structural relationships) models. Examples of the the use of LISREL models of the demand for health include Erbsland, Ried and Ulrich (1995), Hakkinen (1991), van Doorslaer (1987), van de Ven and van der Gaag (1982), Wagstaff (1986, 1993), Wolfe and van der Gaag (1981). van de Ven and Hooijmans (1991) and van Vliet and van Praag (1987) concentrate on the derivation of a health status indexes from the MIMIC model. Behrman and Wolfe (1987) estimate a structural model of health production functions for maternal and child health in Nicaragua; their latent variables include health inputs such as nutrition, along with community and maternal health endowments.

An illustration of the MIMIC approach is van der Gaag and Wolfe's (1991) study which uses data on adults and children from the 1975 Rochester Community Child Health Survey. The problem they address is that health has to be proxied by multiple indicators, none of which are a perfect measure of health. To set the scene for their model they show that principal components analysis can be used to reduce the dimensions of the problem; in their case 26 health indicators are reduced to 4 independent factors. They also show that socio-economic factors affect health, and that the estimated effects depend on the particular measure of health that is used. The relationship between socio-economic factors, desired health, and the demand for medical care is explored through a structural model,

$$H^* = x\alpha + \varepsilon_1 \tag{76}$$

$$D_j = z\beta_{1j} + H^*\beta_{2j} + \varepsilon_{2j} \quad , \quad j=1,..,4 \tag{77}$$

where H* is the (unobserved) desired health status, x and z are socio-economic variables, and the D_js are four observed measures of the demand for medical care. This is combined with the measurement models,

$$HP_l = H^*\gamma_j + \varepsilon_{3l} \quad , \quad l=1,..,L \tag{78}$$

where the HP are proxy measures of health. (76)-(78) are estimated by maximum likelihood as a LISREL model. This assumes joint normality of the error terms and makes use of covariance restrictions to identify the model, so that the unobservable H* is proxied by a linear combination of the health indicators.

van der Gaag and Wolfe (1991) are careful to point out that the kind of health indicators and measures of health care utilisation that commonly arise in survey data are often discrete variables. The normality assumption, used in the LISREL estimation, may not be plausible when dealing with discrete variables.

5.2 *Nonlinear models*

5.2.1 A framework

Blundell and Smith (1993) provide a general framework which is useful to categorise simultaneous equation models involving limited dependent variables. The model consists of an observation mechanism for a limited dependent variable (y_1),

$$y_{1i} = g(y^*_{1i}, y^*_{3i}) \tag{79}$$

and structural equations,

$$y^*_{1i} = \alpha_1 h(y^*_{1i}, y^*_{3i}) + \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i} \tag{80}$$

$$y_{2i} = y^*_{2i} = \alpha_2 h(y^*_{1i}, y^*_{3i}) + x_{2i}\beta_2 + \varepsilon_{2i} \tag{81}$$

The presence of the additional latent variable y^*_{3i} allows for the possibility of sample selection bias. In models without selection bias $g(y^*_{1i}, y^*_{3i}) = g(y^*_{1i})$ and $h(y^*_{1i}, y^*_{3i}) = h(y^*_{1i})$. In many of the applications discussed below y_{1i} is a binary variable and y_{2i} is continuous; and this is the example that will be pursued here.

Blundell and Smith draw attention to the distinction between, what they call, Type I and Type II models. In Type I models $h(y^*_{1i}) = y^*_{1i}$, and the identification condition for the model is $\alpha_1 = 0$. This implies that a Type I specification is appropriate if the structural model is based on a simultaneous equations involving the latent variables,

$$y^*_{1i} = \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i} \tag{82}$$

$$y_{2i} = \alpha_2 y^*_1 + x_{2i}\beta_2 + \varepsilon_{2i} \tag{83}$$

An example of the use of a Type I specification in health economics is Hamilton et al.'s (1997) study of the impact of unemployment on mental health. In their model y^*_1 is a latent index of employability and y_{2i} is mental health, measured by the Psychiatric Symptom Index. However if their structural model had predicted that an individual's actual employment status influenced their mental health, then a Type II specification would be more appropriate.

In a Type II model $h(y^*_{1i}) = y_{1i}$. This is appropriate if the outcome y_{2i} depends on the actual realisation y_{1i} ,

$$y^*_{1i} = \alpha_1 y_{1i}, + \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i} \tag{84}$$

$$y_{2i} = \alpha_2 y_{1i} + x_{2i}\beta_2 + \varepsilon_{2i} \tag{85}$$

Type II specifications raise the problem of coherency conditions. These reflect the logical consistency of the model and are required for the model to have unique reduced form solutions. For example in the model (84) and (85) the restriction, $\alpha_1 + \alpha_2\gamma_1 = 0$, ensures that the probabilities $P(y_{1i} = 0)$ and $P(y_{1i} = 1)$ sum to one.

Estimation of the LDV equations in Type I and Type II models requires different approaches. The Type I specification gives two equations to estimate; one, (82), with the LDV as dependent variable and one, (83), with the continuous dependent variable. Various estimators are available for the LDV equation (82). Of these, two have been favoured in the health economics literature. The two-step or IV estimator replaces the actual values of y_{2i} with fitted values from OLS estimates of the reduced form. The use of predicted values means that the covariance matrix of the estimates should be adjusted to allow for the additional sampling variability (see e.g. Maddala, 1983). The conditional maximum likelihood approach (CML); developed by Smith and Blundell (1986) for the Tobit model and by Rivers and Vuong (1988) for the probit; adds the OLS residuals to the equation. The t-statistic for the residuals provides a simple test for the exogeneity of y_2.

Blundell and Smith (1993) propose an estimator for the Type II LDV equation (84). Again this is based on the CML approach and uses,

$$y^*_{1i} = \gamma_1 \tilde{y}_{2i} + x_{1i}\beta_1 + \rho_1 \tilde{u}_{2i} + \varepsilon_{1i} \qquad (86)$$

where $\tilde{y}_{2i}$ and $\tilde{u}_{2i}$ are obtained from linear IV estimation of (85), using the estimate of α_2 to compute,

$$\tilde{y}_{2i} = y_{2i} - \alpha_2 y_{1i} \qquad (87)$$

This estimator is applied by Sutton and Jones (1997), in a comparison of Type I and Type II specifications of a model of levels and styles of drinking, using data from the British Health and Lifestyle Survey.

5.2.2 Applications

In two related papers, Kenkel (1990, 1991) estimates models for health related behaviour in which continuous measures of health knowledge are treated as endogenous regressors, due to unobservable heterogeneity bias, and replaced by fitted values. Kenkel (1990) uses a survey of 5,336 household from a 1975-76 survey carried out by the Centre for Health Administration Studies and the National Opinion Research Center (CHAS-NORC) of the University of Chicago and looks at the relationship between a general index of health knowledge and physician visits. The probability of a physician visit is modelled using the two-stage probit estimator, replacing the actual values of health knowledge with fitted values from an OLS reduced form. The number of visits is estimated using a simultaneous equation version of the sample selection model, using fitted values along with the Inverse Mill's Ratio from a reduced form probit equation. Kenkel (1991) uses data from the 1985 U.S. National Health Interview Survey for three measures of health related behaviour, smoking, drinking and exercise. These are all censored and Tobit models are used. In all cases the Smith-Blundell test rejects the exogeneity of health knowledge. Kenkel discusses the goodness of fit of the OLS reduced forms. He argues that the results are reasonable for the measures of knowledge about the health effects of smoking and exercise ($R^2 \approx 0.12$-0.19), but rather poor for alcohol ($R^2=0.02$).

Bollen et al. (1995) use data from the Tunisian Demographic and Health Survey to illustrate the practical relevance of Monte Carlo experiments on the performance of different estimators for simultaneous equation probit models reported by Guilkey et al. (1992). Their model involves a binary measure of contraceptive use and a measure of the family's desired number of children; which for the purposes of the analysis is treated as continuous and susceptible to unobservable heterogeneity bias. Like Guilkey et al., they apply a range of estimators: these are the standard probit model, the two-step probit estimator, the conditional ML estimator (CML), FIML, GMM, and a LISREL specification. In the case of the two-step and CML estimators, they rely on Monte Carlo evidence from Guilkey et al. to justify using the standard estimates of the covariance matrix, rather than adjusting the standard errors to allow for the fact that predicted values are being used rather than actual values. Overidentification tests are used to assess the validity of the instruments. These are implemented by comparing the log-likelihood for the model with fitted values and with an unrestricted version in which instruments are added to the equation directly. In their empirical application the exogeneity of the desired number of children cannot be rejected and the simple one-step probit model is favoured. The empirical results are used to reinforce the message from Guilkey et al.'s Monte Carlo evidence; that the performance of the two-step estimators relative to the simple probit model depends on the goodness of fit in the reduced form equations and on the degree of identification, reflected by the number of regressors (x) that are common to both equations.

5.2.3 Switching regressions

The models discussed above include the case of an endogenous binary variable which, in effect, shifts the intercept of the regression function under different regimes. The switching regression model extends this to deal with the case where the whole regression function, slope coefficients as well as the intercept, switches under different regimes. Examples from the health economics literature include O'Donnell's (1993) study of disability and labour supply, in which the income function depends on an individual's labour market status; and Gaynor's (1989) model of nonprice competition within group practices, in which the regression equation for the efficient price locus switches between regimes when demand is constrained or unconstrained.

O'Donnell (1993) uses data from the UK OPCS Disability Survey to investigate the influence of disability benefits on labour market participation by disabled people. The nature of the tax-benefit system means that individuals face a non-convex budget constraint and labour market participation is modelled using a fixed hours specification. A linear utility model leads to a structural labour market participation index,

$$d^*_i = \alpha(y_{1i} - y_{0i}) + x_i\beta + \varepsilon_i \qquad (88)$$

This gives the net utility from working, and depends on the gap between income in work (y_{1i}) and income out of work (y_{0i}), along with socio-economic characteristics x_i. The problem with estimating (88) is that, for a particular individual, only one level of income can be observed. In order to measure the income gap, incomes have to be predicted using reduced form functions,

$$y_{1i} = z_{1i}\alpha_1 + \varepsilon_{1i} \quad , \quad \varepsilon_1 \sim N(0, \sigma^2_1) \tag{89}$$

$$y_{2i} = z_{2i}\alpha_2 + \varepsilon_{2i} \quad , \quad \varepsilon_2 \sim N(0, \sigma^2_2) \tag{90}$$

Because labour market participation is a choice that depends on the levels of y_{1i} and y_{0i}, this gives a switching regression model with endogenous switching. To estimate the model (89) and (90) are substituted into (88) to give a reduced form participation equation. This is estimated as a probit model, and the inverse Mill's ratio is added to the income equations to obtain selectivity corrected estimates. The structural model is identified by exclusion restrictions on β. As long as the model is over-identified, the predicted values of y_{1i} and y_{0i} from the income equations can then be used to obtain consistent, but inefficient, estimates of the parameters of the structural participation equation. In his empirical results, O'Donnell finds that the income gap has a significant effect on labour market participation, although the magnitude of the effect is sensitive to the functional form adopted. He uses the estimates to simulate the impact of the introduction of Disability Working Allowance on employment.

Hay (1991) estimates a variant of the switching regression model with a multinomial logit participation criterion. This allows him to estimate a model for physician's incomes in which their choice of specialty (between GP, internal medicine, and other specialties) may involve self-selection and be influenced by income differentials. Using U.S. data from the Seventh Periodic Survey of Physcians for 1970, he finds the estimated effect of income on choice of specialty changes sign in estimates that take account of selectivity bias.

6. Longitudinal and hierarchical data

6.1 Multilevel models

Multilevel models are used to analyse data that fall naturally into hierarchical structures consisting of multiple macro units, and multiple micro units within each macro unit. Emphasis is placed on defining and exploring variations at each level of the hierarchy after conditioning on the set of explanatory variables of interest. To illustrate the basic structure of a multilevel model consider a simple linear model consisting of two levels which may represent patients (i=1,...,n) nested within hospitals (j=1,....,m). y_{ij} represents the outcome of interest which is related to a vector of explanatory variables x in the following manner:

$$y_{ij} = x_{ij}\beta + \mu_j + \varepsilon_{ij} \tag{91}$$

Assume that the random error term for patient i in hospital j, ε_{ij}, has zero mean and constant variance σ^2_ε. The effects of hospitals are estimated through μ_j which is assumed random and again has a mean of zero and constant variance σ^2_μ. Finally assume that patient and hospital effects are uncorrelated, $\text{cov}(\varepsilon_{ij},\mu_j) = 0$.

For the i-th patient within the j-th hospital, the conditional variance is $var(y_{ij}|x_{ij}\beta) = \sigma^2_\mu + \sigma^2_\varepsilon$ and hence, the overall variance is partitioned into components for both hospitals and patients. The partitioning of the variance in this manner leads to the intra-group correlation coefficient, $\rho = \sigma^2_\mu (\sigma^2_\mu + \sigma^2_\varepsilon)^{-1}$, which measures the strength of nesting within the data hierarchy and is fundamental to the estimation procedures for multilevel models. In the presence of a non-zero intra-group correlation, estimation usually proceeds through the use of generalised least squares (GLS). Various estimation routines have been developed for the analysis of hierarchical data structures and these are reviewed in Rice and Jones (1997).

An alternative to the use of GLS is the Generalised Estimating Equations (GEE) approach; however this is principally concerned with estimates of fixed part parameters (for explanatory variables) rather than exploring the random part. GEE is typically used for clustered data, where there are a large number of clusters. These kinds of data are common in evaluations of prevention programmes which randomise clusters of individuals, rather than specific individuals, to treatment programmes. Norton et al. (1996) use GEE estimates for linear and logistic regressions in an evaluation of Drug Abuse Resistance Education (DARE) using individual data based on a random sample of schools.

It is conceivable that the relationship between an explanatory variable and the response is not the same across all hospitals. Certain hospitals may have the effect of increasing the average response (for example, length of stay) of younger patients whilst decreasing the stays of older patients. The exploration of different 'higher level effects' can be obtained by the inclusion of random coefficients, such that the slope effect associated with an explanatory variable (x_{ij}) can be represented by,

$$y_{ij} = x_{ij}\beta + x_{ij}\gamma_j + \mu_j + \varepsilon_{ij} \qquad (92)$$

In (92) there are three random terms, two of which are random at the hospital level, γ_j and μ_j. This highlights the use of random coefficients by allowing regression coefficients to vary across level 2 units. However more complex variance structures can be introduced at any level of the hierarchy, including level 1, and this may lead to interesting interpretations and better model specification.

The models discussed so far represent the most basic form of a multilevel model where a continuous response is linearly related to a set of explanatory variables and the structure of the hierarchy is simple. In terms of the contributions to health and health economics research, more complex multilevel models may have the most to offer. For example, interest may be focused on the efficiency of both clinicians and provider units when assessing performance. In such a situation the hierarchy consists of patients within clinicians within provider units, and a multilevel model containing three levels is required. Alternatively, data may consist of a series of repeated measurements on patients attending different hospitals. This structure can be modelled using three levels; observations within patients, within hospitals. In reality, clinicians may operate in more than one hospital. In such situations the hierarchy is termed cross-classified.

This occurs when individuals within a lower level cluster are grouped into a different higher level unit than peers from the same cluster.

Many health applications are not suited to a simple model with a linear link function and further extensions to incorporate generalised linear models, including link functions for logit, probit, Poisson, negative binomial, duration (survival) and multinomial models may be specified. The range of applications of multilevel models in health economics is discussed in Rice and Jones (1997).

An example of a linear multilevel model is the analysis of intertemporal preferences for future health by Cairns and van der Pol (1997). Survey respondents were asked to identify what future level of benefit make them indifferent between a specified benefit to be received one year in the future and the more distant delayed benefit. Each respondent was asked to provide estimates of their chosen future level of benefit for two different periods of delay. From the sample data collected implied discount rates for each respondent were calculated and regressed against the set of explanatory variables. The results compare an OLS specification and a multilevel specification of a hyperbolic discounting model. First, it appears that the OLS standard errors are underestimated, and hence the significance of the coefficients are exaggerated. Second, the partitioning of the variance between that observed across responses within respondents and that across respondents themselves allows the intra-class correlation to be estimated. The vast majority of variation (98%) exists across individuals. This suggests that respondents vary greatly in their time preferences, but in comparison appear to be reasonably consistent in applying discount rates to different periods of delay. An advantage of applying a multilevel specification to these data is that the heterogeneity across individuals is modelled whilst preserving degrees of freedom. Due to the lack of multiple responses illicited from individuals, a fixed effects specification would be prohibitive in this application.

Scott and Shiell (1997) apply multilevel analysis with a binary logit link function. Their study analyses the impact of a change in the reimbursement of Australian GPs in 1990. This involved a move from a system based on the length of consultation, to one based on fee descriptors reflecting the content of the consultation, for those GPs on the vocational register. Data is taken from the 1990-91 Australian Morbidity and Treatment Survey. Their working dataset consists of 4,185 consultations for upper respiratory tract infections and sprain/strains, nested within 412 GPs, within 25 types of local area. Three binary dependent variables are investigated measuring prescribing, therapeutic treatments, and counselling. The multilevel model can be expressed in terms of the log-odds ratio for patient i being treated by GP j,

$$\log[\pi_{ij}/(1-\pi_{ij})] = x_{ij}\beta + z_j\gamma + \mu_j + \varepsilon_{ij} \qquad (93)$$

where the x's are measured patient characteristics and the z's are measured GP characteristics. Estimation is based on software which linearises the model and uses a quasi-likelihood procedure. The results do not show a significant effect of the change in reimbursement, proxied by membership of the vocational register, on counselling or treatment, but they do show that prescribing is reduced.

6.2 Random versus fixed effects

The literature on Panel data techniques places emphasis on the relative merits of treating higher level units as random or fixed effects. In model (91), the individual effects (μ_j) are specified as random effects, but they could be specified as fixed effects, to be estimated together with the β's. The choice of specification requires careful consideration and may be determined by the data generating process and the type of inference sought. If individual effects are not of intrinsic importance in themselves, and are assumed to be random draws from a population of individuals and that inferences concerning population effects and their characteristics are sought; then a random specification may be more suitable. However, if inferences are to be confined to the effects in the sample only, and the effects themselves are of substantive interest, then a fixed effects specification may be more appropriate.

Another important consideration is whether the explanatory variables are correlated with the effects. In such circumstances, random or fixed effects approach may lead to very different estimates, and again careful consideration of the model specification is warranted. The situation can be extended to the multilevel model depicted in (91). When μ_j and x_{ij} are correlated, and group sizes are relatively small, the iterative generalised least squares estimator for the parameters β will be inconsistent. Treating the effects μ_j as fixed and applying a least squares dummy variable (LSDV) or within-groups/covariance (CV) estimator leads to consistent estimates. However, when group sizes are large, the two estimators can be shown to be equivalent (see Blundell and Windmeijer (1997)).

In the situation where an explanatory variable is correlated with the higher level effects, and the sole concern of the analyst is the consistent estimation of the parameters associated with the explanatory variables or the mean effect of the higher levels, a fixed effects specification is preferable. However, in the multilevel framework, intrinsic interest lies in the estimation and interpretation of higher level variances, after conditioning on the set of explanatory variables. Rice et al. (1997) develop a conditioned iterative estimator (CIGLS) that attempts to combine the consistency of the fixed effects estimator and the efficiency and estimates of the higher level effects provided by the GLS estimator.

6.3 Fixed effects in panel data

6.3.1 Linear models

Applied work in health economics frequently has to deal with both the existence of unobservable individual effects that are correlated with relevant explanatory variables, and with the need to use nonlinear models to deal with qualitative and limited dependent variables. The combined effect of these two problems creates difficulties for the analysis of longitudinal data; particularly if the model includes dynamic effects such as lagged adjustment or addiction.

To understand these problems, first consider the standard linear panel data regression model, in which there are repeated measurements (t=1,...., T) for a sample of n individuals (i=1,.....,n),

$$y_{it} = x_{it}\beta + \mu_i + \varepsilon_{it} \tag{92}$$

Failure to account for the correlation between the unobservable individual effects (μ) and the regressors (x) will lead to inconsistent estimates of the βs. Adding a dummy variable for each individual will solve the problem, but the least squares dummy variable approach (LSDV) may be prohibitive if there are a large number of cross section observations. The fixed effects can be swept from the equation by transforming variables into deviations from their within-group means. Applying least squares to the transformed equation gives the covariance or within-groups estimator of β (CV). Similarly, the model could be estimated in first differences to eliminate the time-invariant fixed effects.

One disadvantage of using mean deviations or first differences, is that parameters associated with any time invariant regressors, such as gender or years of schooling, are swept from the equation along with the fixed effects. Kerkhofs and Lindeboom (1997) describe a simple two-step procedure for retrieving these parameters; in which estimates of the fixed effects from the differenced equation are regressed on the time invariant variables. This is applied to a model of the impact of labour market status on self-assessed health.

The within-groups estimator breaks-down in dynamic models such as,

$$y_{it} = \alpha y_{it-1} + \mu_i + \varepsilon_{it} \quad , \ \varepsilon_{it} \sim iid \tag{93}$$

This is because the group mean, $y_{it-1} = (1/T)\sum_t y_{it-1}$, is a function of ε_{it} and ε_{it-1}. An alternative is to use the differenced equation,

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \varepsilon_{it} \tag{94}$$

in which case both y_{it-2} and Δy_{it-2} are valid instruments for Δy_{it-1} .

First differences are used by Bishai (1996) to deal with individual and family fixed effects in a model of child health. He develops a model of child health production which emphasises the interaction between a caregiver's education and the amount of time they actually spend caring for the child. The aim is to get around the confounding of, effectively time invariant, levels of education with unobservable (maternal) health endowments. This is done by comparing the productivity of child care time given by members of the family with different levels of education. The model is estimated using the 1978 Intrafamily Food Distribution and Feeding Practices Survey from Bangladesh and the estimator used is the lagged instruments fixed effects estimator (LIFE) of Rosenzweig and Wolpin (1988). This uses differencing to remove the fixed effects, and then estimates the model by 2SLS, using lagged values of childcare time, family resource allocation, and child health as instruments to deal with the potential endogeneity of health inputs and the measures of health.

6.3.2 *The conditional logit estimator*

Now consider a nonlinear model, for example a binary choice model based on the latent variable specification,

$$y^*_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}, \text{ where } y_{it} = 1 \text{ if } y^*_{it} > 0 \tag{95}$$

Then, assuming that the distribution of ε_{it} is symmetric with distribution function F(.),

$$P(y_{it} = 1) = P(\varepsilon_{it} > -x_{it}\beta - \mu_i) = F(x_{it}\beta + \mu_i) \tag{96}$$

This illustrates the "problem of incidental parameters": as $n \to \infty$ the number of parameters to be estimated (β, μ_i) also grows. In linear models β and μ are asymptotically independent, which means that taking mean deviations or differencing allows the derivation of estimators for β that do not depend on μ. In general this is not possible in nonlinear models and the inconsistency of estimates of μ carries over into estimates of β. An exception to this general rule is Chamberlain's conditional logit estimator.

Chamberlain (1980) shows that $\sum_t y_{it}$ is a sufficient statistic for μ_i. This means that conditioning on $\sum_t y_{it}$ allows a consistent estimator for β to be derived. For example with T=2, $\sum_t y_{it} = 0$ is uninformative as it implies that $y_{i1}=0$ and $y_{i2}=0$. Similarly $\sum_t y_{it} = 2$ is uninformative as it implies that $y_{i1}=1$ and $y_{i2}=1$. But there are two ways in which $\sum_t y_{it} = 1$ can occur; either $y_{i1}=1$ and $y_{i2}=0$, or $y_{i1}=0$ and $y_{i2}=1$. Therefore analysis is confined to those individuals whose status changes over the two periods. Using the logistic function,

$$P(y_{it} = 1) = F(x_{it}\beta + \mu_i) = \exp(x_{it}\beta + \mu_i)/(1 + \exp(x_{it}\beta + \mu_i)) \tag{97}$$

it is possible to show that,

$$P[(0,1)|(0,1) \text{ or } (1,0)] = \exp((x_{i2}-x_{i1})\beta)/(1 + \exp((x_{i2}-x_{i1})\beta)) \tag{98}$$

In other words, the standard logit model can be applied to differenced data.

Bjorklund (1985) uses the conditional logit model to analyse the impact of the occurrence and duration of unemployment on mental health using data from the Swedish Level of Living Survey. This includes longitudinal data which allows him to focus on individuals whose mental health status changed during the course of the survey. Bjorklund's estimates compare the conditional logit with cross section models applied to the full sample. He finds that the cross section estimates cannot, on the whole, be rejected when compared to the panel data estimates.

6.3.3 *Parameterising the individual effect*

Another approach to dealing with correlated individual effects is to specify $E(\mu|x)$. For example, in dealing with a random effects probit model Chamberlain (1980,1984) suggests using,

$$\mu_i = x_i\alpha + u_i \quad , \quad u_i \sim \text{iid } N(0, \sigma^2) \tag{99}$$

where $x_i=(x_{i1},....,x_{iT})$. Then the distribution of y_{it} conditional on x but marginal to μ_i has the probit form,

$$P(y_{it}=1) = \Phi[(1+\sigma^2)^{-½}(x_{it}\beta + x_i\alpha)] \tag{100}$$

This could be estimated by maximum likelihood (ML), but Chamberlain suggests a minimum distance estimator.

Labeaga (1993, 1996) develops the Chamberlain approach to deal with situations that combine a dynamic model and limited dependent variables. In Labeaga (1993) he uses panel data from the Spanish Permanent Survey of Consumption, dating from the second quarter of 1977 to the fourth quarter of 1983. Data on real household expenditure on tobacco is used to estimate the Becker-Murphy (1988) rational addiction model; a model that includes past and future consumption as endogenous regressors. The data contain around 40 per cent of zero observations and a limited dependent variable approach is required. The problems of endogeneity and censoring are dealt with separately; using a GMM estimator on the sample of positive observations to deal with endogeneity and using reduced form T-Tobit models to deal with the limited dependent variable problem.

In Labeaga (1996) the two problems are dealt with simultaneously. To illustrate, consider a structural model for the latent variable of interest (say the demand for cigarettes),

$$y^*_{it} = \alpha y^*_{it-1} + x_{it}\beta + x_{it-1}\gamma + z_i\eta + \mu_i + \varepsilon_{it} \tag{101}$$

This allows for dynamics in the latent variable (y^*) and the time varying regressors (x) as well as time invariant regressors (z). The observed dependent variable (y) is related to the latent variable by the observation rule,

$$y^*_{it} = g(y_{it}) \tag{102}$$

where g(.) represents any of the common LDV specifications; such as probit, Tobit, etc..

This specification raises two problems; the inconsistency of ML in nonlinear models with fixed effects and a fixed T, and the correlation between the fixed effect and y^*_{it-1}. Labeaga's solution to this problem combines Chamberlain's approach to correlated individual effects with the within-groups estimator. Assume,

$$\mu_i = w_i\alpha + u_i \text{ and } E(y^*_{i0}|w_i) = w_i\theta \quad (103)$$

where $w_i = [x_{i1}, \ldots, x_{iT}, z_i$, and nonlinear terms in x_i and $z_i]$. Using these assumptions it is possible to derive T reduced form equations, one for each cross section of data,

$$y^*_{it} = w_i\pi_t + e_{it} \quad (104)$$

Each of these can be estimated using the appropriate LDV model, implied by g(.), and specification tests can be carried out on these reduced form models. Once reduced form estimates of π_t have been obtained for each of the cross sections they could be used in a minimum distance estimator. However Labeaga suggests applying the within-groups estimator to equation (101) using the reduced form fitted values of the latent variables (y^*_{it} and y^*_{it-1}). This gives consistent estimates of (α, β, γ), although they are less efficient than the minimum distance estimator. This approach can also deal with continuous endogenous explanatory variables (y_2) by using predictions from the OLS reduced form,

$$E(y_{2it}|w_i) = w_i\pi_2 \quad (105)$$

in the within-groups estimation.

Labeaga's (1993, 1996) results confirm the existence of addiction effects on the demand for cigarettes, even after controlling for unobservable individual heterogeneity. They show evidence of a significant, but inelastic, own-price effect.

López (1997) makes use of Labeaga's approach to estimate the demand for medical care using the Spanish Continuous Family Expenditure Survey. The dependent variable measures expenditure on non-refundable visits to medical practitioners; for which 60 per cent of households make at least one purchase during the 8 quarters that they are measured. This leads López to use an infrequency of purchase specification for the LDV model g(.). He adopts the model of Blundell and Meghir (1987) which allows a separate hurdle for non-participation (identified as no purchases during 8 quarters) and which makes use of the identifying condition that E(y*)=E(y). In specifying the demand for medical care López combines the logarithmic version of the Grossman model with the partial adjustment model used by Wagstaff (1995). The estimates, for the impact of age, education, and the log(wage), show that controlling for censoring and unobservable individual effects does influence the results. This is to be expected, as unobservable heterogeneity is likely to be a particular problem in the use of expenditure survey data which does not contain any direct measures of morbidity.

The work of Dustmann and Windmeijer (1996) brings together many of the ideas discussed so far in this section. They develop a model of the demand for health care based on a variant of the Grossman model in which the demand for health capital is derived solely from the utility of increased longevity. Given the optimal path for health, they assume that there are transitory random shocks to the individual's health. If these fall below a threshold, the individual visits their GP. The model implies that the demand for medical care will depend on the ratio of the initial values of the individual's marginal utilities of wealth and of health; in other words the model contains an unobservable individual effect. The model is estimated with the first four waves of the

German Socio-Economic Panel for 1984-87, using a sample of males who are measured throughout the period and who report visits to a GP. Poisson and negbin2 models are estimated for the number of visits and logit models are estimated for contact probabilities. The specifications of the Poisson and negbin2 models are discussed in more detail below in Section 7.

Dustmann and Windmeijer compare three strategies for dealing with the individual effects. The first is to use a random effects specification. In the negbin2 model the GEE approach is used to allow for the clustering of the data. For the logit model, a nonparametric approach is adopted. This approximates the distribution of unobservable heterogeneity using a finite set of mass points, μ_s, with associated probabilities, p_s (Heckman and Singer, 1984). The likelihood function for this model is,

$$L = \prod_i \sum_s p_s[\prod_t (\lambda_{its})^{yit} (1-\lambda_{its})^{(1-yit)}] \quad (106)$$

where

$$\lambda_{its} = \exp(x_{it}\beta + \mu_s)/(1+ \exp(x_{it}\beta + \mu_s)) \quad (107)$$

and μ_s and p_s are parameters to be estimated. This finite density has been used in other health economics applications, using both count data and survival data, and these are discussed in Sections 7 and 8.

The second strategy is to parameterise the individual effects. They adopt Mundlak's (1978) approach and parameterise the individual effects as a function of the group means for the time varying regressors (they report that they found very similar results with Chamberlain's approach of using all leads and lags of the variables).

The third strategy is to use conditional likelihood estimates of the logit and Poisson models. The log-likelihood for the conditional Poisson is similar to the logit model and takes the form,

$$\text{Log}L = \sum_i \sum_t \Gamma(y_{it} + 1) - \sum_i \sum_t y_{it} \log[\sum_s \exp(-(x_{it} - x_{is})\beta)] \quad (108)$$

where $\Gamma(.)$ is the gamma function ($\Gamma(q) = \int_0^\infty p^{q-1}e^{-p}\, dp$). Overall they find that the second and third strategies, that control for correlated effects, give similar estimates but that they differ dramatically from the random effects specifications. With the fixed effect estimators, the estimated impact of current income is reduced and becomes insignificant. This is consistent with their theoretical model which predicts that permanent rather than transitory income will affect the demand for health, and that the ratio of marginal utilities of wealth and health is a function of lifetime income.

6.3.4 A semiparametric approach: the pantob estimator

The Ministry of Health in British Columbia gives enhanced insurance coverage for prescription drugs to residents aged 65 and over. Grootendorst (1997) uses the "natural experiment" of someone turning 65 to investigate whether the effect of insurance is permanent or transitory, and whether changes are concentrated among those on low incomes. He uses longitudinal claims data for around 18,000 elderly

people for 1985-92. This dataset does not include measures of health status and it has to be treated as an "individual specific fixed endowment subject to a common rate of decay", which is modelled as a fixed effect, μ_i, and an (observable) age effect.

The measure of prescription drug utilisation is censored at the deductible limit and Grootendorst uses Honoré's (1992) panel Tobit estimator (pantob). This estimator deals with censoring and fixed effects, and allows for a non-normal error term. It only requires that the latent variable (y*), after controlling for covariates, is independently and identically distributed for each individual over time. For the case of T=2,

$$y^*_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}, \quad t=1,2 \tag{109}$$

If ε_{i1} and ε_{i2} are i.i.d. then the distribution of (y^*_1, y^*_2) is symmetric around a 45° line through $(x_{i1}\beta, x_{i2}\beta)$. This symmetry gives a pair of orthogonality conditions which imply objective functions that can be used to derive estimators of β. Honoré shows that the estimators are consistent and asymptotically normal for T fixed and $n\to\infty$. Grootendorst's results suggest that there is no permanent effect on drug use, except for low income males. There is little evidence of a transitory effect and it appears that insurance coverage only makes a minor contribution to the growth in utilisation.

7. Count data

7.1 The basic models

Count data regression is appropriate when the dependent variable is a non-negative integer-valued count, y = 0,1,2,..... . Typically these models are applied when the distribution of the dependent variable is skewed to the left, and contains a large proportion of zeros and a long right hand tail. The most common examples in health economics are measures of health care utilisation, such as numbers of GP visits or the number of prescriptions dispensed over a given period.

Cameron and Trivedi (1986) use a range of measures of health care utilisation from the 1977-78 Australian Health Survey (AHS), and this dataset has become a test-bed for many of the recent methodological innovations in the area. Cameron et al. (1988) use a sample of single person households from the AHS and their dependent variables include the number of hospital admissions and the number of days in hospital over the previous year, along with the number of prescribed and the number of non-prescribed medicines taken. Cameron and Trivedi (1993) use the same set of models to illustrate conditional moment tests for independence of the different count variables. Cameron and Windmeijer (1996) use the same data and models as Cameron and Trivedi (1986) to compare a range of models of goodness of fit for count data regressions, favouring those based on deviance residuals. Cameron and Johansson (1997) use the count of visits to (non-doctor) health professionals to illustrate a new estimator based on squared polynomial expansions of the Poisson model. Mullahy (1997b) uses the

measure of number of consultations in the previous two weeks to explore the role of unobservable heterogeneity in accounting for excess zeros in count data.

Other applications to health care utilisation include Cauley (1987), who estimates Poisson regressions for the number of outpatient visits during a year, using a random sample of individuals from the Southern California region of Kaiser Permante Medical Care Programs. Grootendorst (1995) uses self-reported utilisation of medicines by individuals aged 55-75 in the 1990 Ontario Health Survey. Pohlmeier and Ulrich (1995) use cross-section data from the 1985 wave of the German Socio-Economic Panel to estimate hurdle models for the demand for ambulatory care, measured by the number of physician visits during the year. The same dataset is used by Geil et al. (1997), who exploit the unbalanced panel data for 1984-89 and 1992-94 to estimate models for the number of hospital trips each year. Primoff et al. (1995) use data from the 1987 U.S. National Medical Expenditure Survey to estimate a negbin model of mothers' demand for paediatric ambulatory care. Deb and Trivedi (1997) use the same survey to estimate models for six different measures of health care utilisation by the elderly; these include office visits and hospital outpatient visits to both physicians and non-physicians, along with emergency room visits, and inpatient stays. Coulson et al. (1995) use information on the number of prescriptions filled or re-filled over two weeks, among a sample of Medicare enrolled Pennsylvanians. Häkkinen et al. (1996) use information from telephone surveys on physician visits, over the previous six months, to analyse the impact of recession on the use of physician services in Finland. Gurmu et al. (1997) use data from Santa Barbara and Ventura counties taken from the 1986 Medicaid Consumer Survey to estimate models for the number of doctor and health centre visits over a four month (120 day) period, presenting separate results for Medicaid eligible recipients and AFDC beneficiaries. Windmeijer and Santos Silva (1997) and Santos Silva and Windmeijer (1997) use the British Health and Lifestyle Survey to estimate models for the number of GP visits over the past month. Gerdtham (1997) uses measures of the number of physician visits and weeks of care over the past year from the Swedish Level of Living Survey for 1991.

Despite this emphasis on measures of health care utilisation, count data models have proved useful in other areas. Mullahy (1997a) uses data on cigarette smoking from the 1979 U.S. National Health Interview Survey, and on birthweight from the 1988 Child Health Supplement of the NHIS. Kenkel and Terza (1997) use count data on the number of drinks consumed over the two weeks, from the 1990 U.S. National Health Interview Survey.

To understand the nature of count data models consider the following simple example. Assume that the probability of an event (e.g. a GP visit), during a brief period of time (dt), is constant and proportional to its duration. So the probability equals λdt, where λ is known as the intensity of the process. Now consider the count of events from zero up to time t, say (y,t). These are random variables, and the discrete density function must satisfy,

$$f(y,t+dt) = f(y-1,t)\lambda dt + f(y,t)(1-\lambda dt) \qquad (110)$$

Letting $dt \rightarrow 0$ gives a differential equation which solves to give,

$$f(y,t) = e^{-\lambda t}[(\lambda t)^y/y!] \quad (111)$$

which is the joint density of y and t. This yields two additional distributions; the first for the count of events (y) over a fixed interval of time (t=1); and the second for the time (t) until the first occurrence of the event (y=1), or the "time until failure". This illustrates the point that the count data models discussed in this section are, in general, dual to the duration models discussed in Section 8.

Setting t=1, gives the starting point for count data regression; the Poisson process,

$$P(y_i) = e^{-\lambda i}\lambda^{yi} / y_i! \quad (112)$$

This gives the probability of observing a count of y_i events, during a fixed interval. In order to condition the outcome (y) on a set of regressors (x), it is usually assumed that,

$$\lambda_i = E(y_i|x_i) = \exp(x_i\beta) \quad (113)$$

An important feature of the Poisson model is the equidispersion property; that $E(y_i|x_i)=Var(y_i|x_i)=\lambda_i$. Experience shows that this property is often violated in empirical data. In particular, the overwhelming majority of the empirical studies of health care utilisation cited above show evidence of overdispersion ($E(y_i|x_i)<Var(y_i|x_i)$). With overdispersion, the Poisson model will tend to under-predict the actual frequency of zeros, and of values in the right hand tail of the distribution. The need for tests and remedies for overdispersion provide the motivation for many of the methodological developments discussed below.

There are two basic approaches that have been used to estimate count data regressions. Maximum likelihood estimation (ML) uses the fully specified probability distribution and maximises the log-likelihood,

$$LogL = \sum_i \log[P(y_i)] \quad (114)$$

For the Poisson model, the ML estimator solves the first order conditions,

$$x'(y - \lambda) = x'(y - \exp(x\beta)) = 0 \quad (115)$$

If the conditional mean specification is correct but there is under- or overdispersion, then the ML estimates of the standard errors will be biased. However the theory of pseudo-maximum likelihood (PML) estimation ensures that the estimates of β are consistent, and the standard errors can be adjusted by using an appropriate estimator of the covariance matrix (see e.g., Gourieroux et al. (1984), Mullahy (1997c), Windmeijer and Santos Silva (1997)).

The first-order moment condition (115) implies an alternative formulation of the Poisson model, as a nonlinear regression equation,

$$E(y_i|x_i) = \exp(x_i\beta) \quad (116)$$

This is the exponential conditional model (ECM) discussed in Section 4. An alternative approach to estimation, suggested by (116), is to use moment-based estimators, such as nonlinear least squares (NLLS) or generalised method of moments (GMM). For example the GMM estimator minimises,

$$(y - \lambda)'xW^{-1}x'(y - \lambda) \tag{117}$$

where W is a positive definite weighting matrix. As this approach only uses the first moment rather than the full probability distribution, it is more robust than ML. In fact the exponential conditional model encompasses other parametric specifications, such as the geometric and the negative binomial, both of which have the same conditional expectation.

The negative binomial specification allows for overdispersion by specifying, $\exp(x_i\beta+\mu_i)=\exp(x_i\beta)\eta_i$ where η_i is a gamma distributed error term (see e.g., Cameron and Trivedi, 1986). Then

$$P(y_i) = \{\Gamma(y_i+\psi_i)/\Gamma(\psi_i)\Gamma(y_i+1)\}(\psi_i/(\lambda_i+\psi_i))^{\psi_i}(\lambda_i/(\lambda_i+\psi_i))^{y_i} \tag{118}$$

where $\Gamma(.)$ is the gamma function. Letting the "precision parameter" $\psi=(1/a)\lambda^k$, for $a>0$, gives.

$$E(y) = \lambda \text{ and } Var(y) = \lambda + a\lambda^{2-k} \tag{119}$$

This leads to two special cases: setting k=1 gives the negbin 1 model with the variance proportional to the mean, $(1+a)\lambda$; and setting k=0 gives the negbin 2 model where the variance is a quadratic function of the mean, $\lambda + a\lambda^2$. Setting a=0 gives the Poisson model, and this nesting can be tested using a conventional t-test. The negative binomial has been applied extensively in studies of health care utilisation; examples include Cameron and Trivedi (1986), Cameron et al. (1988), Cameron and Windmeijer (1996), Cameron and Johannson (1997), Geil et al. (1997), Gerdtham (1997), Grootendorst (1995), Häkkinen et al. (1996), Pohlmeier and Ulrich (1995).

7.2 Excess zeros

Overdispersion is one source of excess zeros in count data. Mullahy (1997b) emphasises that the presence of excess zeros "is a strict implication of unobserved heterogeneity". In other words, the existence of unobservable heterogeneity may be sufficient to explain excess zeros, without recourse to alternative specifications such as zero inflated or hurdle models. He concentrates on the case where heterogeneity is modelled as a mixture; $\exp(x_i\beta+\mu_i)= \exp(x_i\beta)\eta_i$, with $E(\eta_i)=1$. This includes the negbin model as a special case. Mullahy demonstrates that $P(y_i=0)$ is greater for mixing models than for the Poisson model (where $\eta_i=1$ for all i). A similar result applies to the probability of events in the upper tail of the distribution. The intuition behind these results is that the additional dispersion associated with mixing spreads the distribution

out to the tails. In this sense, the phenomenon of excess zeros is no more than a symptom of overdispersion.

However it may be that there is something special about zero observations *per se*, and an excess of zero counts may not be associated with increased dispersion throughout the distribution. This may reflect the participation decision, and the underlying model of economic behaviour. Many studies of health care utilisation have emphasised the principal-agent relationship between doctor and patient and stressed the distinction between patient initiated decisions, such as the first contact with a GP, and decisions that are influenced by the doctor, such as repeat visits, prescriptions, and referrals (see e.g. Pohlmeier and Ulrich, 1995). There are two approaches which place particular emphasis on the role of zeros; zero inflated models and hurdle, or two-part, models.

The "zero inflated" or "with zeros" model is a mixing specification which adds extra weight to the probability of observing a zero (see e.g., Mullahy, 1986). This can be interpreted as a splitting mechanism which divides individuals into non-users, with probability $q(x_{1i}\beta_1)$, and potential-users, with probability $1\text{-}q(x_{1i}\beta_1)$. So the probability function for the zero inflated Poisson model, $P^{ZIP}(y|x)$ is related to the standard Poisson model, $P^{P}(y|x)$, as follows,

$$P^{ZIP}(y|x) = 1(y=0)q + (1-q)P^{P}(y|x) \tag{120}$$

Zero inflated Poisson and negbin models can be estimated by maximum likelihood. However researchers often report problems in getting the estimates to converge when the full set of regressors are included in the splitting mechanism (see e.g., Grootendorst, 1995, Gerdtham, 1997).

In the count data literature, unlike the limited dependent variable literature discussed in Section 4, hurdle and two-part specifications are treated as synonymous. The hurdle model assumes the participation decision and the positive count are generated by separate probability processes $P_1(.)$ and $P_2(.)$.The log-likelihood for the hurdle model is,

$$\begin{aligned} LogL &= \sum_{y=0} \log[1-P_1(y>0|x)] + \sum_{y>0} \{ \log[P_1(y>0|x)] + \log[P_2(y|x,y>0)]\} \\ &= \{\sum_{y=0} \log[1-P_1(y>0|x)] + \sum_{y>0} \log[P_1(y>0|x)]\} + \{\sum_{y>0} \log[P_2(y|x,y>0)]\} \\ &= LogL_1 + LogL_2 \end{aligned} \tag{121}$$

This shows that the two parts of the model can be estimated separately; with a binary process ($LogL_1$) and the truncated at zero count model ($LogL_2$). Mullahy (1986) introduces the hurdle specification for Poisson and exponential models, while Pohlmeier and Ulrich (1995) extend it by using a negbin 1 specification for both stages. Grootendorst (1995) applies the hurdle model with a probit for the first stage and a negbin 2 model for the second, while Häkkinen et al. (1996) and Gerdtham (1997) use a logit for the first stage and a negbin 2 model for the second stage.

Grootendorst (1995) provides an empirical comparison of hurdle and zero inflated specifications. The study uses data from the 1990 Ontario Health Survey to analyse the

impact of copayments on the utilisation of prescription drugs by the elderly, exploiting the fact that Ontario residents become eligible for zero copayments under the Ontario Drug Benefit Program on their 65th birthday. Zero inflated and hurdle models are not parsimonious, often doubling the number of parameters to be estimated. As always, more complicated models may be prone to over-fitting, and to allow for this Grootendorst uses within-sample forecasting accuracy to evaluate their performance. The models are estimated on a random sample of 70 per cent of the observations. The estimated models are used to compute predictions for the remaining 30 per cent (the forecast sample). Models are then compared on the basis of the mean squared error for the forecast sample. In addition to the split-sample analysis, Voung's non-nested test is computed. The hurdle models outperform the other specifications on all of the criteria. Having established this, Grootendorst goes on to show evidence of heteroscedastity in both the probit and negbin components of the model and parameterises the heterogeneity, but the comparison of models is not repeated.

Pohlmeier and Ulrich (1995) are careful to point out that a limitation of the hurdle model is that it implies that the measure of repeat visits to the doctor relates to a single spell of illness, an issue that may be especially problematic with their annual data. This issue is explored by Santos Silva and Windmeijer (1997) who propose some alternative two stage count models that allow for multiple spells of illness. The observed total number of visits (y) is modelled as,

$$y = \sum_{j=1}^{S} (1 + R_j) \tag{122}$$

where S is the number of illness spells, and R_j is the number of referrals, or repeat visits, during the jth spell. It should be clear that this definition of an illness spell implies that the individual will always make at least one visit to the doctor when they are ill. This perspective leads Santos Silva and Windmeijer to question the need for a truncated model in the second stage of hurdle models. However this seems to be an empirical issue; and allowing individuals to have zero visits during an illness spell may be relevant in studies of unmet need.

(122) implies that y has a stopped sum distribution. Santos Silva and Windmeijer consider two special cases. When S is Poisson and the R_j are independent identical Poisson variates, y has a Thomas distribution. When S is Poisson and $(1+ R_j)$ are logarithmic, y has a negative binomial distribution. The stopped sum specification allows S and R to be parameterised separately, as functions of variables that influence the first visit and that influence referrals. In the light of this, Santos Silva and Windmeijer argue that failure to recognise that the negbin model may reflect a two stage decision process and hence to parameterise the dispersion, may bias comparisons in favour of hurdle models.

Given assumptions about the distributions of S and R_j, the model could be estimated by ML. But pseudo-ML results do not apply, and misspecification of the stopped sum distribution can lead to inconsistent estimates. Instead, Santos Silva and Windmeijer rely on a first-order moment condition and derive a GMM estimator. They use the hypothesis of a single spell (S=1) to generate testable overidentifying restrictions. The estimator is applied to data on the number of GP visits over the past month from the British Health and Lifestyle Survey. They find that the overidentification test does not

reject the hypothesis that the observations are generated by a single spell of illness, suggesting that the hurdle specification may be adequate. The implication is that data collected over longer periods, such as a year, may be prone to the problem of multiple spells, and that, where possible, information should be collected for separate illness spells or episodes of care.

It is often argued that the zero inflated model illustrates the fact that excess zeros can arise even when there is no unobservable heterogeneity (see e.g., Grootendorst, 1995, Mullahy, 1997). For example Grootendorst (1995) argues that comparing the negbin 2 model with a zero inflated negbin 2 allows the analyst to discriminate between unobservable heterogeneity and the splitting mechanism. However a recent paper by Deb and Trivedi (1997) puts a different perspective on the issue. They interpret the zero inflated model as a restrictive special case of a general mixture model with unobservable heterogeneity.

Deb and Trivedi deal with unobservable heterogeneity by using a finite mixture approach. The intuition is that observed counts are sampled from a mixture of different populations. They argue that zero inflated models are a special case of the mixture model, in which the zero counts alone are sampled from a mixture of two populations (non-users and potential users). Their model is implemented using a finite density estimator, where each population, j, is represented by a probability mass point, p_j, (see Heckman and and Singer, 1984). The C-point finite mixture negbin model takes the form,

$$P(y_i|.) = \sum_{j=1}^{C} p_j.P_j(y_i|.), \quad \sum_{j=1}^{C} p_j = 1 \quad , \; 0 \leq p_j \leq 1 \qquad (123)$$

where each of the $P_j(y_i|.)$ is a separate negbin model, and the p_js are estimated along with the other parameters of the model.

The model is applied to the demand for medical care among individuals aged 66 and over, in the 1987 U.S. National Medical Care Expenditure Survey. Demand is measured by six different measures of utilisation for a one year period, and the finite mixture model is compared to hurdle and zero inflated specifications. The finite mixture models are estimated by maximum likelihood, using two and three points of support. The models are compared on the basis of likelihood ratio (LR) and information criterion tests (IC), along with measures of goodness of fit. The negbin 1 models with two points of support are preferred on the basis of these statistical criteria. Deb and Trivedi interpret the points of support as two latent populations of "healthy" and "ill" individuals, reflecting unobserved frailty. Perhaps it is not surprising that a model of health care utilisation among the elderly over a full year which splits the population in this way proves more applicable than the zero inflated and hurdle models, which split individuals into sub-populations of users and non-users.

While Deb and Trivedi apply a finite density estimator to the distribution of unobservable heterogeneity, Gurmu (1997) adopts a semiparametric approach, using a Laguerre series approximation of the unknown density function. This is applied to hurdle models because, unlike the standard model, misspecification of the density leads to inconsistent estimates of the conditional mean in hurdle models. The Laguerre polynomials are complex, but they do have closed form solutions and the model can be

estimated by maximum likelihood. The model nests the Poisson hurdle model and the negbin hurdle with a binary logit for the first stage. In order to balance goodness of fit and parsimony, the number of terms in the Laguerre polynomials is selected according to the Akaike information criterion (AIC = -2Logl - 2(number of free parameters)). Estimates of models for the number of doctor and health centre visits from the 1986 Medicaid Consumer Survey suggest that the semiparametric estimator dominates the Poisson and negbin hurdle models, in terms of the maxmised log-likelihood and the AIC. According to the AIC, a first order polynomial is preferred for the first stage, in other words, a logit model is adequate. While a second order polynomial is preferred at the second stage, giving a specification with greater flexibility than the standard negbin model.

Cameron and Johansson (1997) propose a new estimator that uses squared polynomial expansions around a Poisson baseline. This differs from Gurmu's (1997) approach in that the expansion is around the count density itself, rather than around the density of unobservable heterogeneity. This affects the mean as well as the dispersion. The model is estimated by maximum likelihood using a fast simulated annealing algorithm to deal with the problem of multiple local optima. Cameron and Johansson argue that their estimator is particularly suited for underdispersed data, which is rare in health applications. But for overdispersed data it provides an alternative to the negbin model. They apply the estimator to (non-doctor) health professional visits in the 1977-78 Australian Health Survey and find that their preferred specification, based on a 5th order polynomial, outperforms a negbin 2 model.

7.3 Unobservable heterogeneity and simultaneity biases

Count data models typically assume that unobservable heterogeneity is uncorrelated with the regressors (the same is true of the duration models discussed in Section 8). Mullahy (1997a) argues that this assumption may not hold in many applications, particularly when the unobservable heterogeneity (μ) represents unmeasured omitted regressors. He cites the example of health care utilisation, where μ may reflect an individual's propensity for illness, in which case regressors measuring an individual's insurance coverage may be prone to self selection bias. Similarly, Dustmann and Windmeijer's (1996) model suggests that health care utilisation will depend on correlated individual effects reflecting the ratio of the initial values of the individual's marginal utilities of wealth and of health. The problem may not be confined to individual characteristics; Pohlmeier and Ulrich (1995) argue that unobservable heterogeneity may reflect supply side factors that are nor recorded in individual survey data. These variables may well be correlated with individual characteristics that influence their choice of provider as well as their rate of utilisation of health care. The presence of correlated unobservable heterogeneity means that the standard estimators (ML, PML, NLLS) are inconsistent estimators of β. Mullahy (1997a) proposes the use of nonlinear instrumental variables, estimated by the generalised method of moments (GMM), as a fairly general solution to this problem.

The standard nonlinear instumental variables estimator deals with the case in which unobservables are additively separable,

$$y_i = \exp(x_i\beta) + \mu_i + \varepsilon_i \tag{124}$$

But if μ is to be regarded as an omitted variable it may seem more natural to treat measured and unmeasured regressors "symmetrically" (see e.g., Mullahy, 1997a, and Terza, 1997). This implies that a multiplicative specification should be used, including μ in the linear index,

$$y_i = \exp(x_i\beta + \mu_i) + \varepsilon_i = \exp(x_i\beta)\eta_i + \varepsilon_i \tag{125}$$

While this specification may seem more natural, it raises problems for the use of nonlinear IV estimators. In this context, the assumptions that define a set of valid instruments, z, are,

$$E(y|x,\eta,z) = E(y|x,\eta) \tag{126}$$

$$E(\eta|z) = 1 \tag{127}$$

Now consider the "standard" residual,

$$u_i = y_i - \exp(x_i\beta) \tag{128}$$

where, from (125),

$$u_i = \exp(x_i\beta)(\eta_i - 1) + \varepsilon_i \tag{129}$$

The problem is that this expression involves the product of functions of x and η. So, in general, $E(u|z)\neq 0$, even if (126) and (127) hold. This means that nonlinear IV will be an inconsistent estimator of β. Mullahy's (1997a) solution to this problem is to transform the model so that the transformed residuals (u^T) do satisfy the standard conditions for the consistency of IV. Let,

$$\begin{aligned} u^T_i &= u_i/\lambda_i \\ &= u_i/\exp(x_i\beta) \\ &= \exp(-x_i\beta)y_i - 1 \\ &= \eta_i + \exp(-x_i\beta)\varepsilon_i \end{aligned} \tag{130}$$

The transformed residual is additively separable in η_i, and Mullahy shows that $E(u^T|z)=0$. He then derives an optimal GMM estimator using the transformed residuals to define the moment conditions.

The choice between multiplicative and additive specifications is taken up by Windmeijer and Santos Silva (1997) in the context of simultaneous equations models for count data. They emphasise that, in general, a particular set of instruments, z, will not be orthogonal to both u_i and u^T_i. They appear to be sceptical of the claim that a multiplicative specification is more natural, and argue that the choice is an empirical issue. This can be settled using tests for the overidentifying restrictions in cases where there are more instruments than endogenous regressors.

Windmeijer and Santos Silva use data from the 1991 British Health and Lifestyle Survey to investigate simultaneous equations models for GP visits, in which self-assessed health is treated as a binary endogenous regressor. They adopt the Blundell and Smith (1993) framework, discussed in Section 5, and compare type I and type II specifications. In the type II model, recorded health status is assumed to influence GP visits. In the type I model it is the latent health index that influences the number of visits. The coherency conditions for the type II model imply that the model is only logically consistent when it is specified as a recursive system. In other words, the type II specification can only be coherent when the endogeneity of self-assessed health stems from unobservable heterogeneity bias rather than classical simultaneous equations bias. Additive and multiplicative specifications of the type II model are estimated by GMM (alternative estimators for the multiplicative model are discussed by Terza (1997)); the type I specification is estimated using a two-step approach. The tests of the overidentifying restrictions favour the additive specification, although Hausman tests do not reject the exogeneity of self-assessed health.

8. Survival analysis

8.1 Survival and duration data

Statistical models of "time until failure" have tended to be labelled survival analysis in the epidemiology and biostatistics literature, while the labour economics literature uses the label duration analysis. In health economics, the techniques have been applied to a range of datasets. The most obvious application of survival analysis is to individual lifespan and mortality rates; usually in the context of models of individual health production. For example, Behrman et al. (1990) use the Dorn survey of mortality among U.S. veterans. While Behrman et al. (1991) analyse racial inequality in age specific death rates for males from the U.S. Retirement History Survey (RHS). The RHS is also used by Butler et al. (1989) in a competing risks model for transitions into re-employment or death. Forster and Jones (1997a) use data on mortality from the British Health and Lifestyle Survey (HALS) to estimate a model of the demand for longevity.

However the techniques are not confined to studies of mortality rates. Philipson (1991) uses the child health supplement of the 1991 U.S. National Health Interview Survey (NHIS) to analyse the time elapsed before a child has their first MMR vaccination. Douglas and Hariharan (1994) use the 1978 and 1979 smoking supplements of the NHIS to estimate a model for the age of starting smoking; while Forster and Jones (1997b) use the HALS dataset to analyse the number of years that someone smokes and the decision to quit. Morris et al. (1994) use data from a social experiment involving 36 for-profit nursing homes in San Diego to analyse length of stay by Medicaid recipients. Norton (1995) analyses the time to "spend-down" in nursing homes, modelling the time elapsed before an individual's personal assets are exhausted and they become eligible for Medicaid. Siddiqui (1997) uses the German Socio-

Economic Panel to model the impact of chronic illness and disability on the probability of early retirement using a discrete time hazard rate model. Lindeboom et al. (1995) use a semi-Markov model for sickness, work, and job exit to explain sickness absenteeism among public school teachers in the Netherlands. Bhattachayra et al. (1996) use information on around 440,000 patients from the Japanese Ministry of Health and Welfare's 1990 Patient Survey to estimate a Cox proportional hazards model for the time elapsed between outpatient visits. The delay before adopting a new technology is used by Escarce (1996), to analyse the diffusion of laparoscopic cholecystectomy in a 1992 survey of U.S. surgeons. Hamilton et al. (1996) and Hamilton and Hamilton (1997) use a competing risks specification for post-surgery length of stay and inpatient mortality to estimate the impact of waiting time on surgical outcomes and the volume-outcome relationship.

8.2 Methods

8.2.1 Semiparametric models

The key concept in duration models is the hazard function, defined as the rate of failure at a point in time, given survival to that time. Nonparametric, semiparametric and parametric duration models make assumptions of varying degrees of strength about the hazard function underlying the data generating process. The most commonly used semiparametric duration model is the proportional hazards model of Cox (1972). In this model, the hazard function at time t for individual i, $h_i(t, x_i)$, is defined as the product of a baseline hazard function, $h_o(t)$, and a proportionality factor $\exp(x_i\beta)$,

$$h_i(t, x_i) = h_o(t).\exp(x_i \beta) \tag{131}$$

where x_i is a vector of covariates and β is a parameter vector. The covariates may be time invariant, or the model can be extended to allow for time-varying covariates. For example Philipson (1991) sets out to estimate the "prevalence elasticity" of the demand for MMR vaccinations, and treats regional measles caseloads as a time-varying covariate.

Cox's method is described as being semiparametric because it does not specify the baseline hazard function $h_o(t)$. The partial log-likelihood function for the Cox proportional hazards model is,

$$\text{LogL} = \Sigma_i \ \delta_i\{ x_i\beta - \log(\Sigma_{l\in R_j} \exp(x_l\beta)) \} \tag{132}$$

where δ_i is a dummy variable equal to 1 if the observation exits the process of interest (for example, the age at death of an individual) and 0 if the observation is censored (for example, if an individual is still alive at the end of the data collection period). $l \in R_j$ are those observations in the risk set, R_j, at the time of exit of individual i. R_j includes those observations still alive and uncensored at the time of exit of individual i and whose entry time to the survey is less than or equal to the exit time of the individual

(this controls for left truncation). By conditioning on the risk set the baseline hazard $h_0(t)$ is factored out of the partial likelihood function, in the same way that fixed effects are dealt with in the conditional logit model.

The sampling distribution of the β that maximises the partial likelihood is asymptotically normal, and the standard results of maximum likelihood estimation apply. In the proportional hazards models, the estimates of the parameter vector β measure the effect of a unit change in the covariates of the model on the log of the proportionate shift in the baseline hazard function. Applications of the Cox proportional hazards approach in health economics include Behrman et al. (1990), Behrman et al. (1991), Bhattachyra et al. (1996), Forster and Jones (1997a&b), and Philipson (1995).

A related model, used in Forster and Jones (1997a), is the stratified proportional hazards model,

$$h_{iv}(t) = h_{ov}(t).\exp(x_{iv}\beta) \tag{133}$$

where $h_{iv}(t)$ is the hazard function for individual i in stratum v, β is the common shift parameter vector, x_{iv} is the vector of explanatory variables for individual i in stratum v, and $h_{ov}(t)$.is the baseline hazard function in stratum v. This model can be used when misspecification tests suggests that non-proportional hazards exist for one or more covariates.

8.2.2 Parametric models

Parametric models assume a functional form for the baseline hazard function. Many applied studies compare a variety of different functional forms in order to assess the best empirical specification. Behrman et al. (1990) use the Weibull, log-normal, log-logistic, and generalised gamma. Behrman et al. (1991) use the Weibull and log-logistic. Morris et al. (1994) use the exponential, Weibull, log-normal and generalised gamma. Norton (1995) compares the Weibull, log-normal, log-logistic and generalised gamma. Escarce (1996) uses the Weibull model with and without unobserved gamma heterogeneity.

Specifying the baseline hazard function as $h_o(t)=hpt^{p-1}$ gives the Weibull proportional hazards model,

$$h_i(t) = hpt^{p-1}.\exp(x_i\beta) \tag{134}$$

where p is known as the shape parameter. In the Weibull model, the shape of the baseline hazard function, pt^{p-1}, is shifted by the proportionality factor $h.\exp(x_i\beta)$. The hazard is monotonically increasing for $p>1$, showing increasing duration dependence, and monotonically decreasing for $p<1$, showing decreasing duration dependence. The hazard function, $h(t)=f(t)/S(t)$, can be used to derive the probability density function, f(t), and the survival function, S(t), for the Weibull model, and the likelihood function with right censoring is,

$$L = \prod_i \{f_i(t)/S_i(t)\}^{\delta_i} . S_i(t) \qquad (135)$$

Standard maximum likelihood estimation can be used to obtain estimates of the parameters h, p and β.

The Weibull model may also be estimated in what is called the accelerated time to failure format, which expresses the log of time as a function of the dependent variables and the shape parameter. Taking logs of both sides of (135) and simplifying gives,

$$\log(t_i) = (1/p)\{-\log(h) - x_i\beta + \log(-\log(S_i(t)))\} \qquad (136)$$

where $\log(-\log(S_i(t)))$ has an extreme value distribution. In the accelerated failure time version of the Weibull model, the parameters $-\beta/p$ measure the effect of a one unit change in a covariate on the log of failure time. The Weibull model (and its special case the exponential model, when p=1) is the only parametric model that can be expressed in both the proportional hazards and accelerated time to failure format. But a variety of functional forms are available for the latter. These include non-monotonic hazard functions such as the log-logistic and the generalised gamma.

In their analysis of U.S. data on the age of starting smoking, Douglas and Hariharan (1994) argue that the standard survival analysis may not be appropriate and that a split-population model should be used. The standard survival analysis would treat individuals who had not started smoking by the time of the survey as incomplete spells, and it is assumed that all of these individuals will eventually "fail". The split-population specification allows for the possibility that some people will remain confirmed non-smokers. It augments the standard model by adding a probability, modelled as a probit, that an individual will never fail. A log-logistic specification is used for the hazard function; this is non-monotonic and captures the peak in starting smoking during the mid-teens.

8.2.3 Unobservable heterogeneity

The existence of unobservable heterogeneity will bias estimates of duration dependence. To illustrate, imagine that survival data is sampled from two groups, a "frail" group and a "healthy" group, both of which have constant hazard rates. As time goes by the sample will contain a higher proportion of those with the lower hazard rate; as those with the higher hazard will have died. This will lead to a spurious estimate of negative duration dependence. Kiefer (1988) shows how unobservable heterogeneity can be incorporated by adding a general heterogeneity effect μ and specifying,

$$f(t) = \int f(t|\mu)p(\mu)d\mu \qquad (137)$$

The unknown distribution $p(\mu)$ can be modelled parametrically using mixture distributions. Or a non-parametric approach can be adopted which gives μ a discrete distribution characterised by the mass-points,

$$P(\mu=\mu_i) = p_i \ , \quad i=1,....,I \tag{138}$$

where the parameters (μ_1,....., μ_n,p_1,....., p_n) are estimated as part of the maximum likelihood estimation. This is the basis for the finite support density estimator of Heckman and Singer (1984).

Behrman et al. (1990) and Behrman et al. (1991) provide comprehensive treatments of unobservable heterogeneity in their studies of mortality risks; using parametric, semiparametric, and nonparametric estimators. They adopt two special cases of the Box-Cox conditional hazard used by Heckman and Singer (1984), and they consider two ways in which unobservable frailty (μ) can affect the hazard,

$$h(t|x(t), \mu(t)) = \exp(x(t)\beta + \gamma(t^k-1)/k + \mu(t)) \tag{139}$$

and

$$h(t|x(t), \mu(t)) = \mu(t).\exp(x(t)\beta + \gamma(t^k-1)/k) \tag{140}$$

Their parametric approach uses a normal distribution for $f(\mu)$ in the additive specification and an inverse Gaussian distribution in the multiplicative specification. Both versions of the hazard function can be expressed in the form, $h(t)=h_o(t)\exp(x(t)\beta)$, and their semiparametric estimator uses the Cox partial likelihood approach to factor the baseline hazard out of the likelihood function. The nonparametric Heckman and Singer approach can be applied by using a finite support density estimator for $f(\mu)$.

In addition to these well known approaches, Behrman et al. (1991) apply a maximum penalised likelihood estimator (MPLE). The rationale for this approach is that it avoids over-parameterising the heterogeneity, and it avoids the computational problems associated with the finite density estimator, particularly when there is a high degree of censoring and the distribution of heterogeneity has a long tail. In general, the penalised log-likelihood takes the form,

$$\text{Log}L_\alpha(f) = \sum_i \log(f(x_i)) - \alpha R(f) \tag{141}$$

The penalty term, $\alpha R(f)$, takes account of the "roughness" or local variability in the joint density of the data. The smoothing parameter α, which controls the balance between smoothness and goodness of fit, is typically chosen by cross-validation.

Behrman et al. (1990) evaluate the performance of their models using the maximised value of the likelihood function as a measure of goodness of fit and they test for unobserved heterogeneity using Lancaster's IM test, based on Cox-Snell residuals. They find evidence of heterogeneity but conclude that "modelling of unobserved heterogeneity directly in a proportional hazard setting may not be as important as allowing the covariates to affect the hazard in the highly nonlinear way that the gamma accelerated failure-time model allows". Behrman et al. (1991) find that the "introduction of nonparametric or parametric heterogeneity yields a small improvement in fit, similar parameter estimates, and changed significance levels".

8.3 Competing risks and multiple spells

So far the focus has been on duration models with a single destination; such as an individual's death. But the techniques can be extended to allow for multiple destinations; or competing risks. For example, Butler, Anderson, and Burkhauser (1989) use a competing risks specification with transitions out of retirement either back into employment or due to death. While, in their study of sickness absence among Dutch teachers, Lindeboom et al. (1995) use a three state Markov model that allows transitions from spells of work into sickness absence or exit from the job, and from spells of sickness back into work or exit from the job. Their model uses a partial likelihood approach to allow for school specific fixed effects.

Hamilton and Hamilton's (1997) study of the surgical volume-outcome relationships for patients undergoing surgery for hip fractures in Quebec between 1991-93 provides an example that combines competing risks, unobservable heterogeneity, and fixed effects. They use longitudinal data from the MED-ÉCHO database of hospital discharge abstracts. This allows them to attribute differences in the quality of providers to hospital specific fixed effects, modelled by dummy variables, and to analyse the within-hospital volume-outcome relationship; thereby discriminating between the "practice makes perfect effect" and "selective referral effect" (that hospitals with good outcomes will get more referrals).

Their competing risks specification allows for a correlation between the two outcomes; post-surgery length of stay and inpatient mortality. This is important as, ceteris paribus, a death in hospital is more likely for a patient with a longer length of stay. With two exhaustive and mutually exclusive destinations for discharges, alive (a) or dead (d), the probability of exit to state r, after a length of stay m, for patient i, in hospital h, at period t, is assumed to be,

$$f_r(m_{iht}|x_{iht}) = \lambda_r(m_{iht}|x_{iht}) \prod_{j\in a,d} \exp[-\int_0^{m_{iht}} \lambda_j(u|x_{iht})\, du] \;,\; r=a,b \qquad (142)$$

The first term on the right hand side, $\lambda_r(m_{iht}|x_{iht})$, is the transition intensity; the equivalent of the hazard rate in single destination models. The second term is the survivor function; giving the probability of surviving to m without death or discharge. A proportional hazards specification is used,

$$\lambda_r(.) = \exp(x_{iht}\beta_r + \theta_{hr} + \pi_r\mu)\lambda_{or}(m_{iht}),\; r=a,d \;,\; \text{and } \pi_a=1 \qquad (143)$$

where θ_{hr} is the hospital fixed effect, and a log-logistic baseline hazard is used,

$$\lambda_{or}(m) = (\rho_r\alpha_r m^{\alpha r-1})/(1 + \rho_r m^{\alpha r}) \;,\; \alpha_r>0,\, \rho_r>0 \qquad (144)$$

Unobserved fraility is modelled as the scalar random variable μ, and its distribution is estimated using the Heckman-Singer approach. The likelihood takes the form,

$$L = \prod_i \sum_k p_k.f_a(m_{iht}|x_{iht}, \theta_h, \mu_k)^{\delta ia}.f_d(m_{iht}|x_{iht}, \theta_h, \mu_k)^{\delta id} \;,\; \sum_k p_k=1 \qquad (145)$$

where the points of support (μ_k) and associated probabilities (p_k) are estimated along with the other parameters. Hamilton and Hamilton (1997) use three mass points, which they interpret as a distribution made up of three types of patient.

The results of the study show that when hospital fixed effects are added to the model the coefficient on volume, measured by the logarithm of live discharges, declines substantially and is insignificant. Volume does not have a significant effect on inpatient deaths with or without hospital fixed effects; although cruder models without unobservable heterogeneity and with fewer controls for comorbidities do show a significant effect.

9. Stochastic frontiers

9.1 Cost function studies

A recent systematic review by Aletras (1996) identifies approximately 100 studies which provide evidence on the existence of economies of scale and scope in hospitals. Many of these are econometric studies which use regression analysis to explore the average cost of hospital treatment. Other methods include data envelopment analysis (DEA), market survival methods, and before-and-after studies. These attempts to estimate empirical production functions and cost functions for hospitals and other health care organisations face some common methodological problems.

It would be desirable to define a hospital's output in terms of health outcomes, measured as health gains, but typically these kinds of data are not available and measures of throughput have to be used (e.g., admissions, discharges, number of procedures performed). Output is multi-dimensional, and it is important to control for case-mix, by including variables for the proportion of patients in each specialty, the number of discharges, or the average length of stay by specialty or case-mix grouping. However these case-mix adjustments may miss intra-category variations in severity, and inter-hospital variation in case-mix. Measures of the quantity of output may neglect differences in quality across hospitals, which may bias estimates of economies of scale. Similar arguments apply to the neglect of differences in the quality of inputs. Also, in econometric studies, the level of output is usually assumed to be exogenous, reflecting the demand for health care from patients or purchasers. The possibility of an incomplete agency relationship between purchasers and providers may lead to simultaneous equations bias.

9.2 Frontier models

9.2.1 Cross section estimators

Rather than discussing hospital cost studies in general, this section concentrates on the econometric techniques that have been used to analyse the efficiency of health care organisations, and in particular the use of stochastic frontier models. This builds on earlier surveys by Wagstaff (1989a&b) and Aletras (1996).

Feldstein's (1967) pioneering econometric analysis of hospitals costs in the British NHS uses the following empirical specification,

$$y_i = \beta_o + \Sigma_j \beta_j x_{ij} + u_i \qquad (146)$$

where y_i is the average cost per case and x_{ij} is the proportion of hospital i's patients in the jth case-mix category. In this model the residuals are distributed symmetrically around the cost function and it cannot be interpreted as a frontier. This is relaxed by deterministic cost frontier (DCF) models, which assume $u_i \geq 0$ for all i. In this case the error term moves hospitals above the (deterministic) cost frontier. One estimator for this model is corrected OLS, which simply adjusts the OLS estimates of the intercept β_o and the residuals by adding $\min(u_i)$ to the intercept and subtracting it from the residuals. The drawbacks of this method are that it treats the most efficient hospital as 100 per cent efficient, and that the whole of the error term is assumed to reflect inefficiency. This ignores random "noise" due to measurement errors and unobservable heterogeneity.

To relax these assumptions stochastic cost frontiers (SCF) are based on the two-error model,

$$y_i = \beta_o + \Sigma_j \beta_j x_{ij} + u_i + \varepsilon_i, \quad u_i \geq 0 \qquad (147)$$

where u_i measures inefficiency and ε_i is a random error term. To estimate parametric versions of this model by maximum likelihood it is necessary to make assumptions about the distributions of u and ε. For example Aigner et al. (1977) assume that ε is normal and u is half-normal. Other common assumptions are that u is truncated normal, exponential, or gamma distributed.

Vitaliano and Toren (1994) apply stochastic frontiers to estimate cost inefficiency in New York nursing homes, using the 1987 and 1990 waves of a panel dataset. After experimenting with truncated normal and exponential distributions, they choose to estimate the model using a half normal inefficiency term. They use Jondrow et al.'s method to decompose the estimated error term; this computes an estimate of inefficiency conditional on the estimated residual, $E(u_i | u_i+\varepsilon_i)$. Their results suggest a mean inefficiency of 29 per cent.

Stochastic frontiers are applied to a multiproduct hospital cost function by Zuckerman et al. (1994). They use data on 1,600 U.S. hospitals from the AHA Annual Survey, Medicare hospital cost reports, and MEDPAR data system to estimate translog cost functions that include measures of illness severity, output quality, and patient outcomes. The SCF models are estimated by ML using a half-normal distribution for inefficiency, these suggest a mean inefficiency of 13.6 per cent. The authors are concerned about possible endogeneity of the output measures, and find that Hausman-Wu tests reject exogeneity in non-frontier specifications. However, they are not able to find estimates that converge when instrumental variables are used in the frontier models.

The use of stochastic frontiers is not confined to estimates of hospital cost functions. Gaynor and Pauly (1990) use production frontiers to investigate the effects of different compensation arrangements on productive efficiency in medical group practices. They compare "traditional" production functions, which only include inputs, with "behavioural" functions, which include variables that reflect incentives. Data on 6,353 physicians within 957 group practices, from a survey carried out by Mathematica Policy Research in 1978, are used to estimate stochastic frontiers using normal and truncated normal error components. The potential endogeneity of variables that measure the firm's compensation structure is dealt with using instrumental variables. The results suggest that incentives do influence productivity, with larger groups reducing productivity and greater average experience within a group increasing productivity.

Jones et al. (1997) apply frontier models to individual health production functions; using data from the British Health and Lifestyle Survey to estimate the impact of cigarette smoking on respiratory health. Half-normal, truncated normal, and exponential stochastic frontier models are used to estimate the efficiency with which individuals produce respiratory health, measured by their forced expiratory volume. Instrumental variables are used to deal with the possible confounding effects of unobservable heterogeneity bias. The results show that smoking has a detrimental effect on respiratory health and they identify the specific effects of smoking intensity, duration, and recovery after quitting.

Most cross-section frontier models are estimated by maximum likelihood, imposing specific parametric distributions on both u and ε . Kopp and Mullahy (1990, 1993) propose semiparametric estimators which relax the distributional assumptions about ε, simply requiring that it is symmetrically distributed. Given the symmetry assumption, they are able to derive restrictions for the higher order moments of the composite error term. In Kopp and Mullahy (1990) these moment conditions are used to motivate a GMM estimator, and in Kopp and Mullahy (1993) they are used to motivate a COLS or corrected moment (CM) estimator. These estimators do not seem to have been applied to health data as yet.

9.2.2 *Panel data estimators*

The fact that cross-section models rely on skewness to identify inefficiency is often criticised (see e.g. Wagstaff, 1989b, Skinner, 1994). The danger is that skewness in the distribution of the random error term could be mistakenly attributed to inefficiency. The alternative is to use panel data estimators. On the assumption that inefficiency remains constant over time, the stochastic frontier model takes on a form similar to the standard panel data regression (see equation (92)),

$$y_{i\,t} = \beta_o + \Sigma_j \beta_j x_{ij\,t} + u_i + \varepsilon_{it}, \quad u_i \geq 0 \tag{148}$$

This model can be estimated using fixed or random effects estimators, and the results are subject to the strengths and weaknesses of these estimators, as discussed in Section 6. In particular, the fixed effects models raises the problem of separately identifying inefficiency and the effects of time invariant regressors, while the random effects specification is biased if the inefficiency is correlated with the regressors.

Wagstaff (1989b) uses data on 49 Spanish public hospitals to compare cross section and panel data estimators. Cross section estimates based on the half-normal model suggest that mean cost inefficiency is only 10 per cent, and it is not possible to reject the null hypothesis that there is no skewness. However estimates of the fixed effects specification suggest that around one third of the variation in costs can be attributed to inefficiency. Also the stochastic frontier leads to quite different efficiency rankings than the fixed effects and deterministic cost frontier models. This ambiguity leads Wagstaff to recommend that a range of methods are compared to assess the sensitivity of the efficiency estimates to model specification.

Standard panel data methods do not make use of the fact that inefficiency, u_i, is expected to be non-negative. Koop et al. (1997) acknowledge this and develop a Bayesian fixed effects estimator, using the prior that the inefficiency effects will be one-sided and independent. They also develop a random effects estimator that allows the inefficiency to depend on time invariant hospital characteristics. These estimators are applied to a panel of 382 U.S. non-teaching hospitals for 1987-91. Estimates of a translog cost function show that for-profit hospitals are less efficient, although these results are based on highly aggregated measures of output and may neglect differences in quality.

The assumption that inefficiency remains constant over time can be relaxed. For example, Battese and Coelli (1992) propose a panel data estimator model in which firm specific inefficiency takes the form,

$$u_{it} = \exp\{-\eta(t-T)\}u_i \geq 0 \tag{149}$$

This allows inefficiency to change over time, but on the assumption that the rate of change, η, is common to all firms. The model is estimated by ML, on the assumption that that ε is normal and u is truncated or half-normal. Battese and Coelli (1995) propose an alternative specification in which,

$$u_{it} = z_{it}\delta + \omega_{it} \geq 0 \qquad (150)$$

The z_{it} variables are determinants of cost inefficiency and the distribution of u_{it} is assumed to be truncated normal. Linna (1997) applies both of these models to Finnish panel data covering 43 acute hospitals for 1988-94.

10. Conclusion

In documenting the influence of econometrics on the development of health economics, Newhouse (1987) grouped imports from econometrics under four headings: specification tests, robust estimators, replication, and experimentation. Ten years on, the first two of these remain dominant themes in applied work. Examples of good practice in health econometrics make extensive use of tests for misspecification and explicit model selection criteria. Robust and distribution-free estimators are of increasing importance, and this chapter has given examples of nonparametric, and semiparametric estimators applied to sample selection, simultaneous equations, count data, and survival models. As the use of these techniques widens, it will be interesting to see whether they have an impact on the economic and policy relevance of the results produced. Even if the impact proves to be small, researchers will have greater confidence in results generated by less robust methods.

Published replications of empirical results remain relatively rare, perhaps reflecting the incentives surrounding academic publication in economics. One way in which this deficit may be remedied is through the appearance of more systematic reviews of econometric studies, such as the work of Aletras (1996). The chapter has shown that certain datasets are widely used, allowing results to be compared across studies, and many of the studies reviewed here are careful to compare new techniques with established methods. The use of experimental data remains an exception and most applied studies continue to rely on observational data from secondary sources. However applied work in health economics is likely to be influenced by the debate concerning the use of instrumental variables to analyse social experiments (see e.g. Angrist et al., 1996, Heckman, 1997).

This chapter has illustrated the impressive diversity of applied econometric work over the past decade. It has emphasised the range of models and estimators that have been applied, but that should not imply a neglect of the need for sound economic theory and careful data collection and analysis in producing worthwhile econometric research. Most of the studies reviewed here use individual level data and this has lead to the use of a wide range of nonlinear models, including qualitative and limited dependent variables, along with count, survival and frontier models. Because of the widespread use of observational data, particular attention has gone into dealing with problems of self-selection and heterogeneity bias. This is likely to continue in the future, with the emphasis on robust estimators applied to longitudinal and other complex datasets.

ACKNOWLEDGEMENTS

The text of this chapter draws on joint work with Vassilios Aletras, Paul Contoyannis, Alan Duncan, Martin Forster, Rob Manning, Nigel Rice, Matt Sutton, and Steven Yen.

I am grateful for valuable suggestions and comments from Ignacio Abasolo, Tom Buchmueller, Marty Gaynor, Antonio Giuffrida, Michael Grossman, Don Kenkel, Will Manning, John Mullahy, Edward Norton, Owen O'Donnell, Carol Propper, Belinda South, Joe Terza, and John Wildman.

References

Ahn, H. and J.L. Powell (1993), "Semiparametric estimation of censored selection models with a nonparametric selection mechanism", *Journal of Econometrics, Annals* **58**: 3-30.

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977), "Formulation and estimation of stochastic production function models", *Journal of Econometrics* **6**: 21-37.

Alderson, C. (1997), "Exporing a modified 'fair innings' appraoch to addressing social class inequalities in lifetime health", unpublished M.Sc. dissertation, *University of York.*

Aletras, V. (1996), "Concentration and choice in the provision of hospital services. Technical Appendix 2", *NHS Centre for Reviews and Dissemination, University of York.*

Angrist, J.D. (1995), "Conditioning on the probability of selection to control selection bias", *NBER Technical Working Paper #181.*

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of causal effects using instrumental variables", *Journal of the American Statistical Association* **91**: 444-455.

Atkinson, A.B., J. Gomulka, and N.H.Stern (1984), "Household expenditure on tobacco 1970-1980: evidence from the Family Expenditure Survey", *London School of Economics*, Discussion Paper no.60.

Auster, R., I. Leveson and D. Sarachek (1969), "The production of health an exploratory study", *Journal of Human Resources* **15**: 411-436.

Battese, G.E. and T.J. Coelli (1992), "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India", *Journal of Productivity Analysis* **3**: 153-169.

Battese, G.E. and T.J. Coelli (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data", *Empirical Ecconomics* **20**: 325-332.

Becker, G.S. and K.M. Murphy (1988), "A theory of rational addiction", *Journal of Political Economy* **96**: 675-700.

Behrman, J.R., R.C. Sickles and P. Taubman (1990), "Age-specific death rates with tobacco smoking and occupational activity: sensitivity to sample length, functional form, and unobserved frailty", *Demography* **27**: 267-284.

Behrman, J.R., R. Sickles, P. Taubman and A. Yazbeck (1991), "Black-white mortality inequalities", *Journal of Econometrics* **50**: 183-203.

Behrman, J.R. and B.L. Wolfe (1987), "How does mother's schooling affect family health, nutrition, medical care usage, and household sanitation?", *Journal of Econometrics* **36**: 185-204.

Bhattacharya, J., W.B. Vogt, A. Yoshikawa and T. Nakahara (1996), "The utilization of outpatient medical services in Japan", *Journal of Human Resources* **31**: 450-476.

Bishai, D.M. (1996), "Quality time: how parents' schooling affects child health through its interaction with childcare time in Bangladesh", *Health Economics* **5**: 383-407.

Björklund, A. (1985), "Unemployment and mental health: some evidence from panel data", *Journal of Human Resources* **20**: *469-483.*

Blaylock, J.R. and W.N. Blisard (1992), "Self-evaluated health status and smoking behaviour", *Applied Economics* **24**: 429-435.

Blaylock, J.R. and W.N. Blisard (1993), "Wine consumption by US men", *Applied Economics* **25**: 645-651.

Blundell, R.W. and R.J. Smith (1993), "Simultaneous microeconometric models with censored or qualitative dependent variables", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11.* (Elsevier, Amsterdam) 117-143.

Blundell, R.W. and F.A.G. Windmeijer (1997), "Correlated cluster effects and simultaneity in multilevel models", *Health Economics* **6:** 439-443.

Bolduc, D., G. Lacroix and C. Muller (1996), "The choice of medical providers in rural Bénin: a comparison of discrete choice models", *Journal of Health Economics* **15**: 477-498.

Bollen, K.A., D.K. Guilkey and T.A. Mroz (1995), "Binary outcomes and endogenous explanatory variables: tests and solutions with an application to the demand for contraceptive use in Tunisia", *Demography* **32**: 111-131.

Buchmueller, T.C. and P.J. Feldstein (1997), "The effect of price on switching among health plans", *Journal of Health Economics* **16**: 129-260.

Butler, J.S., K.H. Anderson and R.V. Burkhauser (1989), "Work and health after retirement: a competing risks model with semiparametric unobserved heterogeneity", *The Review of Economics and Statistics* ..: 46-53.

Cairns, J.A. and M. van der Pol (1997), "Saving future lives: a comparison of three discounting models", *Health Economics* **6**: 341-350.

Cameron, A.C. and P. Johansson (1997), "Count data regression using series expansions: with applications", *Journal of Applied Econometrics* **12**: 203-223.

Cameron, A.C. and F.A.G. Windmeijer (1996), "R-squared measures for count data regression models with applications to health care utilization", *Journal of Business and Economic Statistics* **14**: 209-220.

Cameron, A.C. and P.K. Trivedi (1986), "Econometric models based on count data: comparisons and applications of some estimators and tests", *Journal of Applied Econometrics* **1**: 29-53.

Cameron, A.C., P.K. Trivedi, F. Milne and J.Piggott (1988), "A microeconometric model of demand for health care and health insurance in Australia", *Review of Economic Studies* **55**: 85-106.

Cameron, A.C. and P.K. Trivedi (1993), "Tests of independence in parametric models with applications and illustrations", *Journal of Business and Economic Statistics* **11**: 29-43.

Cauley, S.D. (1987), "The time price of medical care", *Review of Economics and Statistics*

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 M. Intrilligator, eds., *Handbook of Econometrics* (North-Holland, Amsterdam): 1247-1318.

Coulson, N.E., J.V. Terza, C.A. Neslusan and B.C. Stuart (1995), "Estimating the moral-hazard effect of supplemental medical insurance in the demand for prescription drugs by the elderly", *AEA Papers and Proceedings* **85**: 122-126.

Deb, P. and P.K. Trivedi (1997), "Demand for medical care by the elderly: a finite mixture approach", *Journal of Applied Econometrics* **12**: 313-336.

Dor, A., P. Gertler, and J. van der Gaag (1987), "Non-price rationing and the choice of medical care providers in rural Cote D'Ivoire", *Journal of Health Economics* **6**: 291-304.

Douglas, S. and G. Hariharan (1994), "The hazard of starting smoking: estimates from a split population duration model", *Journal of Health Economics* **13**: 213-230.

Duan, N. (1983), "Smearing estimate: a nonparametric retransformation method", *Journal of the American Statistical Association* **78**: 605-610.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse (1983), " A comparison of alternative models for the demand for medical care", *Journal of Business and Economic Statistics* **1**: 115-126.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse (1984), " Choosing between the sample-selection and multi-part model", *Journal of Business and Economic Statistics* **2**: 283-289.

Duncan, A.S. and A.M. Jones (1992), "NP-REG: an interactive package for kernel density estimation and nonparametric regression", *Institute for Fiscal Studies*, Working Paper W92/7.

Dustmann and F.A.G. Windmeijer (1996), "Health, wealth and individual effects - a panel data analysis", presented at *Fifth European Workshop on Econometrics and Health Economics.*

Ellis, R.P., D.K. McInnes, and E.H.Stephenson (1994), "Inpatient and outpatient health care demand in Cairo, Egypt", *Health Economics* **3**: 183-200.

Erbsland, M., W. Ried and V. Ulrich (1995), "Health, health care, and the enviroment. Econometric evidence from German micro data", *Health Economics* **4**: 169-182.

Escarcé, J.J. (1996), "Externalities in hospitals and physicina adoption of a new surgical technology: an exploratory analysis", *Journal of Health Economics* **15**: 715-734.

Feldman, R., M. Finch, B. Dowd, and S. Cassou (1989), "The demand for employment-based health insurance plans", *Journal of Human Resources* **24**: 115-142.

Feldstein, M.S. (1967), *Economic analysis for health service efficiency: econometric studies of the British National Health Service.* (North-Holland, Amsterdam).

Forster and Jones (1997a), "Inequalities in optimal life-span: a theoretical and empirical investigation", mimeo, University of York.

Forster and Jones (1997b), "The optimal time path of consumption of an unhealthy good: a theoretical and empirical investigation of smoking durations" , mimeo, University of York.

van der Gaag, J. and B.L. Wolfe (1991), "Estimating demand for medical care: health as a critical factor for adults and children", In G. Duru and J.H.P. Paelinck, eds., *Econometrics of Health Care*, (Kluwer, Amsterdam) 31-58.

Garcia, J. and J.M Labeaga (199.), "A cross-section model with zeros: an application to the demand for tobacco", *Oxford Bulletin of Economics and Statistics*

Gaynor, M. (1989), "Competition within the firm: theory plus some evidence from medical group practice", *RAND Journal of Economics* **20**: 59-76.

Gaynor, M. and M.V. Pauly (1990), "Compensation and productive efficiency in partnerships: evidence from medical group practice", *Journal of Political Economy* **98**: 544-573.

Geil, P., A. Million, R. Rotte and K.F. Zimmermann (1997), "Economic incentives and hospitalization in Germany", *Journal of Applied Econometrics* **12**: 295-311.

Gerdtham, U-G. (1997), "Equity in health care utilization: further tests based on hurdle models and Swedish micro data", *Health Economics* **6**: 303-319.

Gertler, P., L. Locay and W.Sanderson (1987), "Are user fees regressive? The welfare implications of health care financing proposals in Peru", *Journal of Econometrics* **36**: 67-88.

Gourieroux, C.A., A. Monfort, and A. Trognon (1984), "Pseudo maximum likelihood methods: applications to Poisson models", *Econometrica* **52**, 701-720.

Grootendorst, P.V. (1995), "A comparison of alternative models of prescription drug utilization", *Health Economics* **4**: 183-198.

Grootendorst, P.V. (1997), "Health care policy evaluation using longitudinal insurance claims data: an application of the panel Tobit estimator", *Health Economics* **6**: 365-382.

Guilkey, D.K. , T.A. Mroz, and L. Taylor (1992), "Estimation and testing in simultaneous equations models with discrete outcomes using cross section data", unpublished manuscript.

Gurmu, S. (1997), "Semi-parametric estimation of hurdle regression models with an application to medicaid utilization", *Journal of Applied Econometrics* **12**: 225-242.

Haas-Wilson, D., A. Cheadle and R. Scheffler (1988), "Demand for mental health services: an episode of treatment approach", *Southern Economic Journal* **55**: 219-232.

Haas-Wilson, D. and E. Savoca (1990), "Quality and provider choice: a multinomial logit-least squares model with selectivity", *Health Services Research* **2:**, 791-809.

Hakkinen, U. (1991), "The production of health and the demand for health care in Finland", *Social Science and Medicine* **33**: 225-237.

Hakkinen, U., G. Rosenquist and S. Aro (1996), "Economic depression and the use of physician services in Finland", *Health Economics* **5**: 421-434.

Hamilton, B.H. and V.H. Hamilton (1997), "Estimating surgical volume-outcome relationships applying survival models: accounting for frailty and hospital fixed effects", *Health Economics* **6**: 383-395.

Hamilton, B.H., V.H. Hamilton, and N.E. Mayo (1996), "What are the costs of queueing for hip fracture surgery in Canada?", *Journal of Health Economics* **15**: 161-185.

Hamilton, V.H., P. Merrigan and E. Dufresne (1997), "Down and out: estimating the relationship between mental health and unemployment", *Health Economics* **6**: 397-406.

Hay, J.W. (1991), "Phyisicians' specialty choice and specialty income", In G. Duru and J.H.P. Paelinck, eds., *Econometrics of Health Care*, (Kluwer, Amsterdam) 95-113.

Hay, J. and R.J. Olsen (1984), "Let them eat cake: a note on comparing alternative models of the demand for health care", *Journal of Business and Economic Statistics* **2**: 279-282.

Heckman, J.J. (1979), "Sample selection bias as a specification error", *Econometrica* **47:** 153-161.

Heckman, J.J. (1996), "Randomization as an instrumental variable", *Review of Economics and Statistics ..,* 336-341.

Heckman, J.J. (1997), "Instrumental variables. A study of implicit behavioral assumptions used in making program evaluations", *Journal of Human Resources* **32**: 441-461

Heckman, J.J. and B. Singer (1984), "A method of minimizing the distributional impact in econometric models for duration data", *Econometrica* **52**: 271-230.

Honoré, B.E. (1992), "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects", *Econometrica* **60**: 533-565.

Hughes, M.D. (1988), "A stochastic frontier cost function for residential child care provision", *Journal of Applied Econometrics* **3**: 203-214.

Hunt-McCool, J., B.F. Kiker and Y.C. Ng (1994), "Estimates of the demand for medical care under different functional forms", *Journal of Applied Econometrics* **9**: 201-218.

Ichimura, H. and L.F.Lee (1991), "Semiparametric estimation of multiple index models:single equation estimation", in Barnett, W.A., J. Powell and G. Tauchen, eds. *Nonparametric and semiparametric methods in econometrics and statistics* (Cambridge University Press, New York).

Imbens, G.W. and J.D. Angrist (1994), "Identification of local average treatment effects", *Econometrica* **62**: 467-475.

Jones, A.M. (1989), "A double-hurdle model of cigarette consumption", *Journal of Applied Econometrics* **4**: 23-39.

Jones, A.M. (1993), "Starters, quitters and smokers: choice or addiction", prepared for the Inaugural Labelle Lectureship, *CHEPA*, McMaster University.

Jones, A.M., R. Manning and M. Sutton (1997), "The impact of cigarette smoking upon the efficient production of respiratory health", *Centre for Health Economics Technical Paper 5.*

Jonsson, B. and U. Gerdtham (1998), "Healthcare systems internationally compared", in Newhouse, J.P. and A.J. Culyer, eds., *Handbook of Health Economics* (North-Holland, Amsterdam).

Kenkel, D.S. (1990), "Consumer health information and the demand for medical care", *The Review of Economics and Statistics .., 587-595.*

Kenkel, D.S. (1991), "Health behaviour, health knowledge and schooling", *Journal of Political Economy* **99**: 287-305.

Kenkel, D.S. (1995), "Should you eat breakfast? Estimates from health production functions", *Health Ecconomics* **4**: 15-29

Kenkel, D.S. and J.V. Terza (1993), "A partial observability probit model of medical demand", mimeo.

Kerkhofs, M. and M. Lindeboom (1995), "Subjective health measures and state dependent reporting errors", *Health Economics* **4**: 221-235.

Kerkhofs, M. and M. Lindeboom (1997), "Age related health dynamics and changes in labour market status", *Health Economics* **6**: 407-423.

Kiefer, N. (1988), "Economic duration data and hazard functions", *Journal of Economic Literature* **26**: 646-679.

Koop, G., J. Osiewalski and M.F.J Steel (1997), "Bayesian efficiency analysis through individual effects: hospital cost frontiers", *Journal of Econometrics* **76**: 77-105.

Kopp, R.J. and J. Mullahy (1990), "Moment-based estimation and testing of stochastic frontier models", *Journal of Econometrics* **46**: 165-183.

Kopp, R.J. and J. Mullahy (1993), "Least squares estimation of econometric frontier models: consistent estimation and inference", *Scandanavian Journal of Economics* **95**: 125-132.

Labeaga, J.M. (1993), "Individual behaviour and tobacco consumption: a panel data approach", *Health Economics* **2**: 103-112.

Labeaga, J.M. (1996), "A dynamic panel data model with limited dependent variables: an application to the demand for tobacco", mimeo.

Lee L-F., M.R. Rosenzweig and M.M. Pitt (1997), "The effects of improved nutrition, sanitation, and water quality on child health in high mortality populations", *Journal of Econometrics* **77**: 209-235.

Leibowitz, A., W.G. Manning and J.P. Newhouse (1985), "The demand for prescription drugs as a function of cost-sharing", *Social Science and Medicine* **21**: 1063-1069.

Leung, S.F. and S. Yu (1996), "On the choice between sample selection and two-part models", *Journal of Econometrics* **72**: 197-229.

Lewit, E.M., D. Coate and M. Grossman (1981), "The effects of government regulation on teenage smoking", *Journal of Law and Economics* **24**: 545-570.

Lindeboom, M., M. Kerkhofs and L. Aarts (1995), "Sickness absenteeism of primary school teachers in the Netherlands", mimeo, Leiden University.

Linna, M. "Measuring the hospital cost efficiency with panel data models", Paper presented at Sixth European Workshop on Econometrics and Health Economics, Lisbon.

Lopez, A. (1997), "Unobserved heterogeneity and censoring in the demand for health care", unpublished manuscript.

McGuire, A. D.Parkin, D.Hughes and K. Gerard (1993), "Econometric analyses of national health expenditures: can positive economics help to answer normative questions?", *Health Economics* **2**: 113-126.

Maddala, G.S. (1983), *Limited-dependent and qualitative variables in econometrics,* (Cambridge University Press, Cambridge).

Maddala, G.S. (1985), "A survey of the literature on selectivity bias as it pertains to health care markets", In R.M. Scheffler and L.F. Rossiter, eds., *Advances in Health Economics and Health Services Research, Volume 6* (JAI Press, Greenwich Connecticut) 3-17.

Manning, W.G., L. Blumberg and L.H. Moulton (1995), "The demand for alcohol: the differential response to price", *Journal of Health Economics* **14**: 123-148.

Manning, W.G., N. Duan and W.H. Rogers (1987), "Monte Carlo evidence on the choice between sample selection and two-part models", *Journal of Econometrics* **35**: 59-82.

Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler, A. Leibowitz and M.S. Marquis (1987), "Health insurance and the demand for medical care: evidence from a randomized experiment", *American Economic Review* **77**: 251-277.

Manski, C.F. (1993), "The selection problem in econometrics and statistics", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11.* (Elsevier, Amsterdam) 73-84.

Morris, C.N., E.C. Norton and X.H. Zhou (1994), "Parametric duration analysis of nursing home usage", In N. Lange et al., eds., *Case Studies in Biometry*, (John Wiley & Sons Ine, New York) 231-248.

Mullahy, J. (1986), "Specification and testing of some modified count data models", *Journal of Econometrics* **33**: 341-365.

Mullahy, J. (1997a), "Instrumental variable estimation of count data models. Applications to models of cigarette smoking behaviour", *Review of Economics and Statistics*.....

Mullahy, J. (1997b), "Heterogeneity, excess zeros, and the structure of count data models", *Journal of Applied Econometrics* **12**: 337-350.

Mullahy, J. (1997c), "Much ado about two: reconsidering the two-part model in health econometrics", mimeo.

Mullahy, J. and W. Manning (1996), "Statistical issues in cost-effectiveness analysis", in Sloan, F.A. ed., *Valuing health care* (Cambridge University Press, Cambridge) 149-184.

Mullahy, J. and P.R. Portney (1990), "Air pollution, cigarette smoking, and the production of respiratory health", *Journal of Health Economics* **9**: 193-205.

Mullahy, J. and J. Sindelar (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.

Mwabu, G., M. Ainsworth, and A. Nyamete (1993), "Quality of medical care and choice of medical treatment in Kenya. An empirical analysis", *Journal of Human Resources* **28**: 838-862.

Newhouse, J.P. (1987), "Health economics and econometrics", *American Economic Review* **77**: 269-274.

Newhouse, J.P., C.E. Phelps and M.S.M. Marquis (1980), "On having your cake and eating it too. Econometric problems in estimating the demand for health services", *Journal of Econometrics* **13**: 365-390.

Norton, E.C. (1995), "Elderly assets, Medicaid policy, and spend-down in nursing homes", *Review of Income and Wealth* **41**: 309-329.

Norton, E.C., G.S. Bieler, S.T. Ennett and G.A Zarkin (1996), "Analysis of prevention program effectiveness with clustered data using generalized estimating equations", *Journal of Consulting and Clinical Psychology* **64**: 919-926.

O'Donnell, O. (1993), "Income transfers and the labour market participation of disabled individuals in the UK", *Health Economics* **2**: 139-148.

Philipson, T. (1996), "Private vaccination and public health: an empirical examination for U.S. measles", *Journal of Human Resources* **31**: 611-630.

Pitt, M.M. (1997), "Estimating the determinants of child health when fertility and mortality are selective", *Journal of Human Resources* **32**: 129-158.

Pohlmeier, W. and V. Ulrich (1995), "An econometric model of the two-part decisionmaking process in the demand for health care", *Journal of Human Resources* **30**: 339-360.

Primoff, .. Vistnes, J. and V. Hamilton (1995), "The time and monetary costs of outpatient care for children", *American Economic Review Papers and Proceedings* **85**: 117-121.

Rice, N. and A.M. Jones (1997), "Multilevel models and health economics", *Health Economics* **6**: 561-575.

Rice, N., A.M. Jones and H. Goldstein (1997), "Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator", (CIGLS), mimeo.

Rosenbaum, P.R. and D.B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika* **70**: 41-55.

Rosenzweig, M.R. and T.P. Schultz (1983), "Estimating a household production function: heterogeneity, the demand for health inputs, and their effects on birth weight", *Journal of Political Economy* **91**: 723-746.

Rosenzweig, M.R. and K.T. Wolpin (1995), "Sisters, siblings, and mothers: the effect of teenage childbearing and birth outcomes in a dynamic family context", *Econometrica* **63**: 303-326.

Santos Silva, J.M.C. and F.A.G. Windmeijer (1997), "Stopped-sum models for health care demand", presented at *Sixth European Workshop on Econometrics and Health Economics,* Lisbon.

Scott, A. and A. Shiell (1997), "Do fee descriptors influence treatment choices in general practice? A multilevel discrete choice model", *Journal of Health Economics* **16**: 323-342.

Siddiqui, S. (1997), "The impact of health on retirement behaviour: empirical evidence from West Germany", *Health Economics* **6**: 425-438.

Skinner, J. (1994), "What do stochastic cost frontiers tell us about inefficiency?", *Journal of Health Economics* **13**: 323-328.

Stern, S. (1996), "Semiparametric estimates of the supply and demand effects of disability on labor force participation", *Journal of Econometrics* **71**: 49-70.

Sutton, M. and C. Godfrey (1995), "A grouped data regression approach to estimating economic and social influences on individual drinking behaviour", *Health Economics* **4**: 237-247.

Sutton, M. and A.M. Jones (1997), "Levels and styles of drinking: a LDV simultaneous equations approach", mimeo, Univeristy of York.

Terza, J.V. (1997), "Estimating count data models with endogenous switching: sample selection and endogenous treatment effects", *Journal of Econometrics,* forthcoming.

Ullah, A. (1988), "Non-parametric extimation of econometric functionals", *Canadian Journal of Economics* **21**: 625-658.

van Doorslaer, E.K.A. (1987), *Health, knowledge and the demand for medical care.* (van Gorcum, Assen/Maasricht).

van der Gaag, J. and B. Wolfe (1991), "Estimating demand for medical care: health as a critical factor for adults and children", In G. Duru and J.H.P. Paelinck, eds., *Econometrics of Health Care*, (Kluwer, Amsterdam) 31-58.

van de Ven, W.P.M.M. and J. van der Gaag, (1982), "Health as an unobservable: a MIMIC model for health care demand", *Journal of Health Economics* **1**: 157-183.

van de Ven, W.P.M.M. and E.M. Hooijmans (1991), "The MIMIC health status index", In G. Duru and J.H.P. Paelinck, eds., *Econometrics of Health Care*, (Kluwer, Amsterdam) 19-29.

van de Ven, W.P.M.M. and B.M.S. van Praag (1981), "The demand for deductibles in private health insurance",*Journal of Econometrics* **17**: 229-252.

van Vliet, R.C.J.A. and B.M.S. van Praag (1987), "Health status estimation on the basis of MIMIC health care models", *Journal of Health Economics* **6**: 27-42.

Vitaliano, D.F. and M. Toren (1994), "Cost and efficiency in nursing homes: a stochastic frontier approach", *Journal of Health Economics* **13**: 281-300.

Wagstaff, A. (1986), "The demand for health. Some new empirical evidence", *Journal of Health Economics* **5**: 195-233.

Wagstaff, A. (1989a), "Econometric studies in health economics", *Journal of Health Economics* **8**: 1-51.

Wagstaff, A. (1989b), "Estimating efficiency in the hospital sector: a comparison of three statistical cost frontiers", *Applied Economics* **21**: 659-672.

Wagstaff, A. (1993), "The demand for health: an empirical reformulation of the Grossman model", *Health Economics* **2**: 189-198.

Wasserman, J. W.G. Manning, J.P. Newhouse and J.D. Winkler (1991), "The effects of excise taxes and regulations on cigarette smoking", *Journal of Health Economics* **10**: 43-64.

Windmeijer, F.A.G. and J.M.C. Santos Silva (1997), "Endogeneity in count data models; an application to demand for health care", *Journal of Applied Econometrics* **12**: 281-294.

Wolfe, B. and J. van der Gaag, J. (1981), "A new health status index for children", In J. van der Gaag and M. Perlman, eds., *Health, economics, and health economics*, (North-Holland Amsterdam) 283-304.

Yen, S.T. and A.M. Jones (1996), "Individual cigarette consumption and addiction: a flexible limited dependent variable approach", *Health Economics* **5**: 105-117.

Zimmerman Murphy, M. (1987), "The importance of sample selection bias in the extimation of medical care demand equations", *Eastern Economic Journal* **13**: 19-29.

Zuckerman, S., J. Hadley, and L. Iezzoni (1994), "Measuring hospital efficiency with frontier cost functions", *Journal of Health Economics* **13**: 255-280.