# HEALTH ECONOMETRICS: Dealing with the selection problem and individual heterogeneity in observational data.

**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 the Health Economists' Study Group, January 2000**.

This paper is based on extracts from Jones, A.M. (2000) "Health Econometrics", *North-Holland Handbook in Health Economics*, J.P. Newhouse and A.J. Culyer (eds.), Elsevier, forthcoming.
The full text can be downloaded from:

**http://www.york.ac.uk/res/herc/yshe**

# 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 Jones (2000) shows that the range and volume of applied econometric work in health economics has increased dramatically over the past few years. What is more, the prevalence of latent variables, unobservable heterogeneity, and nonlinear models make health economics a particularly rich area for applied econometrics.

# 2. Identification and estimation

## *2.1 The evaluation problem*

The evaluation problem is whether it is possible to 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.

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", for example whether the individual has health insurance or not, on the outcome. The causal effect of interest is,

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

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, $\mu$ represents unobserved covariates, and $\varepsilon$ is a random error term reflecting sampling variability. The problem for inference arises if x and $\mu$ 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 may achieve the desired orthogonality of measured covariates (x) and unobservables ($\mu$); 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.

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 (see e.g., McClellan and Newhouse, 1997). In practice the validity of the IV approach relies on finding appropriate instruments (see e.g., Bound et al., 1995).

iii) Control function approaches to selection bias - these range from parametric methods such as the Heckit estimator to more recent semiparametric estimators (see e.g., Heckman, 1979, Vella, 1998).

### *2.3 Model specification and estimation*

So far, the discussion has concentrated on the evaluation problem and selection bias. More generally, most econometric work in health economics focuses on the problem of finding an approporiate stochastic model to fit the available data. Estimation of regression functions like equation (3) typically requires assumptions about the appropriate conditional distribution for the dependent variable and for the functional relationship with one or more covariates. Failure of these assumptions may mean that estimators lose their desired properties and give biased, inconsistent, or inefficient estimates. For this reason attention should be paid to tests for misspecification and robust methods of estimation.

Classical regression analysis assumes that the regression function is linear and that the random error term has a normal distribution,

$$y_i = x_i\beta + \varepsilon_i \quad , \ \varepsilon_i \sim N(0,\sigma^2) \tag{4}$$

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.

In health economics empirical analysis is complicated further by the fact that the theoretical models often involve inherently unobservable (latent) concepts such as 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; 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; multinomial responses, such as the choice of provider; limited dependent variables, such as expenditure on primary care services, which is censored at zero; integer counts, such as the number of GP visits; or measures of duration, such as the time elapsed between visits.

Maximum likelihood (ML) estimation is widely used in health economics, particularly for nonlinear models involving qualitative or limited dependent variables. ML has desirable properties, such as consistency and asymptotic normality, but these rely on the model being fully and correctly specified. Pseudo (or quasi) maximum likelihood (PML) methods share the properties of ML without having to maintain the assumption that the model is correctly specified (see e.g., Gourieroux et al., 1984, Gourieroux and Monfort, 1993). For the class of distributions belonging to the linear exponential family, which includes the binomial, normal, gamma, and Poisson, the PML estimator of the conditional mean is consistent and asymptotically normal. This means that the conditional mean has to be specified but the conditional variance does not. The main use of PML methods in health economics has been in the context of count data regressions.

Many of the estimators used in health economics fall within the unifying framework of generalised method of moments (GMM) estimation (see e.g., Hall, 1993). This replaces population moment conditions, such as,

$$E[f(x,\beta)] = 0 \tag{5}$$

with their sample analogues,

$$m(\beta) = n^{-1}\sum_i f(x_i,\beta) = 0 \tag{6}$$

The GMM estimator minimises a quadratic form,

$$Q(\beta) = m(\beta)'W\, m(\beta) \qquad (7)$$

where W is a positive definite matrix, and the optimal W can be selected to give asymptotic efficiency. GMM encompasses many standard estimators. For example OLS uses the moment conditions $E[x(y-x\beta)]=0$, instrumental variables with a set of instruments z uses $E[z(y-x\beta)]=0$, and pseudo maximum likelihood uses $E[\partial LnL/\partial\beta)]=0$. Applications of GMM in health economics are found in the context of instrumental variable estimation and count data models.

Quantile regression is another semiparametric method which assumes a parametric specification for the qth quantile of the conditional distribution of y,

$$\text{Quantile}_q\,(y_i|\, x_i) \;=\; x_i\beta_q \qquad (8)$$

but leaves the error term unspecified (see e.g., Buchinsky, 1998). Quantile regression has been applied by Manning et al. (1995) 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.

Many of the estimators discussed in Jones (2000) rely on the approximation provided by asymptotic theory for their statistical properties. Recent years have seen increasing use of bootstrap methods to deal with cases where the asymptotic theory is intractable or where the asymptotics are known but finite sample properties of an estimator are unknown (see e.g., Jeong and Maddala, 1993). The aim of these methods is to reduce bias and to provide more reliable confidence intervals. Bootstrap data are constructed by repeated re-sampling of the estimation data using random sampling with replacement. The bootstrap samples are then used to approximate the sampling distribution of the estimator being used. For example Nanda (1998) uses bootstrap methods to compute standard errors for two stage instrumental variable estimates in a model of the impact of credit programs on the demand for health care among women in rural Bangladesh.

The growing popularity of bootstrap methods reflects the increased availability of computing power. The same can be said for simulation methods. Monte Carlo simulation techniques can be used to deal with the computational intractability of nonlinear models, such as the multinomial probit, which involve higher order intergrals (see e.g., Hajivassiliou, 1993). Popular methods include the GHK simulator and Gibbs sampling. These methods can be applied to simulate sample moments, scores, or likelihood functions. Simulation estimators for the multinomial probit and for simultaneous equation limited dependent variable models have been used in health economics (see e.g., Bolduc et al., 1996, Pudney and Shields, 1997, Hamilton, 1998).

# 3. The selection problem

## *3.1 Manski bounds*

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 the "outcome" y is only observed if the "treatment" d equals 1. Interest centres on the unconditional probability,

$$P(y|x) = P(y|x,d=1)P(d=1|x) + P(y|d=0,x)P(d=0|x) \quad (9)$$

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 known is,

$$P(y|x) \in [\, P(y|x,d=1)P(d=1|x) + \gamma\, P(d=0|x)\,,\ \gamma \in \Gamma\,] \quad (10)$$

where $\Gamma$ is the space of all probability measures on y. To address this problem the statistical literature often assumes independence or ignorable non-response,

$$P(y|x) = P(y|d=0,x) = P(y|d=1,x) \quad (11)$$

This is a strong assumption which asserts that those individuals who do not receive the treatment would respond in the same way to those who do, conditional on the covariates. 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 the censored sampling process. These identifying restrictions relate to functional forms, including exclusion restrictions on the regressors that enter the regression equations, and assumptions about the distribution of the error terms. In the econometrics literature, the traditional approach has been the parametric Heckit model, but recent years have seen the development of less restrictive semiparametric estmators which relax some, but not all, of the identifying restrictions. These are discussed in the next section.

The selection problem is fatal for inferences concerning E(y|x) without identifying restrictions, but Manski shows that it is possible to put bounds on other features of the distribution. This leads to nonparametric estimation of the bounds and to estimators for quantile regressions.

### *3.2 Semiparametric estimators*

The biostatistics literature has seen the development of the propensity score approach, to deal with the problem of identifying treatment effects when there is self-selection bias in the assignment of patients to treatments. In the econometrics literature, this idea is connected to the development of semiparametric estimators for the sample selection model, some of which have been applied in health economics. These estimators focus on relaxing the distributional assumptions about the error terms in the sample selection model and, in particular, they seek to avoid the assumption of joint normality which is required to identify the Heckit model.

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 re-capping a general version of the sample selection model. Assume that the following is observed,

$$y = [x_2\beta_2 + \varepsilon_2]\ \mathbf{1}[\varepsilon_1 > -\psi(x_1)] \tag{12}$$

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,

$$E[y|x,d=1] = x_2\beta_2 + E[\varepsilon_2|x, \varepsilon_1 > -\psi(x_1)] \tag{13}$$

If the distribution of $(\varepsilon_1, \varepsilon_2)$ is independent of $x_1$ and $x_2$, the conditional expectation of $\varepsilon_2$ depends only on $\psi(x_1)$. The propensity score is defined as follows,

$$P(x_1) = P(d=1|x_1) = P[\varepsilon_1 > -\psi(x_1)] \tag{14}$$

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

$$E[y|x,d=1] = x_2\beta_2 + \tau[P(x_1)] \tag{15}$$

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,

$$y_i - y_j \approx [x_{2i} - x_{2j}]\ \beta_2 + \varepsilon_{ij} \tag{16}$$

This leads Ahn and Powell to suggest a weighted IV estimator for $\beta_2$, using kernel estimates of $(P_i-P_j)$ as weights. They 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,

$$E[y|x,d=1) = x_2\beta_2 + \lambda(x_1\beta_1) \tag{17}$$

(15) and (17) both have the partially linear form discussed in section 2.3 and can be estimated using Robinson's (1988) approach.

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,

$$y = z_o + \psi(z_1, z_2) + \varepsilon \tag{18}$$

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,

$$(1/n) \sum [(y - z_0) - E(y-z_0|z_1,z_2)]^2 \tag{19}$$

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

$$\begin{aligned} E(y-z_0|z_1,z_2)] = (1/n-1) [\textstyle\sum_{j\neq i}( y_j - z_{0j})K[(z_{1i} - z_{1j})/h_1, (z_{2i} - z_{2j})/h_2] \\ /(1/n-1) \textstyle\sum_{j\neq i} K[(z_{1i} - z_{1j})/h_1, (z_{2i} - z_{2j})/h_2] \end{aligned} \tag{20}$$

and where K[.,.] is a kernel function and the h's are bandwidths. The Ichimura and Lee 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; $\psi_1$ , $\psi_2 \geq 0$.

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. Longitudinal data

## *4.1 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),

$$y_{it} = x_{it}\beta + \mu_i + \varepsilon_{it} \qquad (21)$$

Failure to account for the correlation between the unobservable individual effects ($\mu$) and the regressors (x) will lead to inconsistent estimates of the $\beta$s. 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 $\beta$ (CV). Similarly, the model could be estimated in first differences to eliminate the time-invariant fixed effects. It should be clear that identification of $\beta$ rests on there being sufficient variation within groups. In practice, fixed effects may only work well when there are many observations and much variation within groups.

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,

$$y_{it} = \alpha y_{it-1} + \mu_i + \varepsilon_{it}, \ \varepsilon_{it} \sim iid \quad (22)$$

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

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \varepsilon_{it} \quad (23)$$

in which case both $y_{it-2}$ and $\Delta y_{it-2}$ are valid instruments for $\Delta y_{it-1}$ as long as the error term ($\varepsilon_{it}$) does not exhibit autocorrelation.

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.

### *4.2 The conditional logit estimator*

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

$$y^*_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}, \ \text{where } y_{it} = 1 \text{ if } y^*_{it} > 0, \ 0 \text{ otherwise.} \quad (24)$$

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

$$P(y_{it} = 1) = P(\varepsilon_{it} > -x_{it}\beta - \mu_i) = F(x_{it}\beta + \mu_i) \quad (25)$$

This illustrates the "problem of incidental parameters": as $n \to \infty$ the number of parameters to be estimated ($\beta$, $\mu_i$) also grows. In linear models $\beta$ and $\mu$ are asymptotically independent, which means that taking mean deviations or differencing allows the derivation of estimators for $\beta$ that do not depend on $\mu$. In general this is not possible in nonlinear models and the inconsistency of estimates of $\mu$ carries over into estimates of $\beta$.

An exception to this general rule is the conditional logit estimator. The conditional logit estimator uses the fact that $\sum_t y_{it}$ is a sufficient statistic for $\mu_i$ (see e.g., Chamberlain (1980)). This means that conditioning on $\sum_t y_{it}$ allows a consistent estimator for $\beta$ 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,

$$P(y_{it} = 1) = F(x_{it}\beta + \mu_i) = \exp(x_{it}\beta + \mu_i)/(1+ \exp(x_{it}\beta + \mu_i)) \qquad (26)$$

it is possible to show that,

$$P[(0,1)|(0,1) \text{ or } (1,0)] = \exp((x_{i2}-x_{i1})\beta)/(1+ \exp((x_{i2}-x_{i1})\beta)) \qquad (27)$$

In other words, the standard logit model can be applied to differenced data and the individual effect is swept out.

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.

### *4.3 Parameterising the individual effect*

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

$$\mu_i = x_i\alpha + u_i \quad , \quad u_i \sim \text{iid } N(0, \sigma^2) \qquad (28)$$

where $x_i=( x_{i1},....,x_{iT})$. Then, by substituting (28) into (24), the distribution of $y_{it}$ conditional on x but marginal to $\mu_i$ has the probit form,

$$P(y_{it} = 1) = \Phi[(1+\sigma^2)^{-½}(x_{it}\beta + x_i\alpha)] \qquad (29)$$

The model could be estimated directly by maximum likelihood (ML), but Chamberlain suggests a minimum distance estimator. This takes the estimates from reduced form probits on $x_i$, for each cross section, and imposes the restrictions implied by (29) to retrieve the parameters of interest $(\beta,\sigma)$.

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, by 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),

$$y^*_{it} = \alpha y^*_{it-1} + x_{it}\beta + x_{it-1}\gamma + z_i\eta + \mu_i + \varepsilon_{it} \quad (30)$$

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,

$$y^*_{it} = g(y_{it}) \quad (31)$$

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,

$$\mu_i = w_i\alpha + u_i \text{ and } E(y^*_{i0}|w_i) = w_i\theta \quad (32)$$

where $w_i$ = [$x_{i1}$,....., $x_{iT}$, $z_i$, and nonlinear terms in $x_i$ and $z_i$]. The second assumption addresses the problem of the initial condition for the value of y*. Using these assumptions it is possible to derive T reduced form equations, one for each cross section of data,

$$y^*_{it} = w_i\pi_t + e_{it} \quad (33)$$

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 $\pi_t$ 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 (30) using the reduced form fitted values of the latent variables ($y^*_{it}$ and $y^*_{it-1}$). This gives consistent estimates of ($\alpha$, $\beta$, $\gamma$), 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,

$$E(y_{2it}|w_i) = w_i\pi_2 \quad (34)$$

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 (1998) 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 a model 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 (1993). 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 do 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.

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, $\mu_s$, with associated probabilities, $p_s$ (Heckman and Singer, 1984). The likelihood function for this model is,

$$L = \prod_i \sum_s p_s[\prod_t (\lambda_{its})^{yit} (1-\lambda_{its})^{(1-yit)}] \quad (35)$$

where

$$\lambda_{its} = \exp(x_{it}\beta + \mu_s)/(1+ \exp(x_{it}\beta + \mu_s)) \quad (36)$$

and $\mu_s$ and $p_s$ are parameters to be estimated. This finite density estimator has been used in other health economics applications, using both count data and survival data.

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,

$$\text{LogL} = \sum_i \sum_t \Gamma(y_{it} + 1) - \sum_i \sum_t y_{it} \log[\sum_s \exp(-(x_{it} - x_{is})\beta)] \qquad (37)$$

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, which 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.

### *4.3 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 on prescription drug use 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, $\mu_i$, 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 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,

$$y^*_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}, \quad t=1,2 \qquad (38)$$

If $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are i.i.d. then the distribution of $(y^*_1, y^*_2)$ is symmetric around a $45^o$ 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 $\beta$. 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.

## 5. 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 be able to place more confidence in earlier results that were 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). This 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).

Jones (2000) illustrates 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 led 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. Also it has benefited from the suggestions made by Will Manning and Edward Norton in their detailed discussions of an earlier version.

I am grateful for valuable suggestions and comments from Ignacio Abasolo, Tom Buchmueller, Tony Culyer, Eddy van Doorslaer, Eric French, Marty Gaynor, Antonio Giuffrida, Julie Glanville, Paul Grootendorst, Michael Grossman, Don Kenkel, Sylvia Kuo, Angel López, John Mullahy, Christophe Muller, Joe Newhouse, Owen O'Donnell, Carol Propper, João Santos Silva, Chris Seplaki, Jim Sears, Harri Sintonen, Frank Sloan, Belinda South, Andrew Street, Joe Terza, Volker Ulrich, Dave Vaness, John Wildman, and Amy Wolaver.

# 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.

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.

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

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.*

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.

Bound, J., D. Jaeger, and R. Baker (1995), "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak", *Journal of the American Statistical Association* **90**: 443-450.

Buchinsky, M. (1998), "Recent advances in quantile regression models", *Journal of Human Resources* **33**: 88-126.

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.

Dustmann, C. 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*, Barcelona.

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

Gourieroux, C.A. and A. Monfort (1993), "Pseudo-likelihood methods", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11. (*Elsevier, Amsterdam) 335-362.

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.

Hajivassiliou, V.A. (1993), "Simulation estimation methods for limited dependent variable models", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11. (*Elsevier, Amsterdam) 519-543.

Hall, A. (1993), "Some aspects of generalized method of moments estimation", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11. (*Elsevier, Amsterdam) 393-417.

Hamilton, B. (1998), "The impact of HMOs on Medicare costs: Bayesian MCMC estimation of a robust panel data Tobit model with survival", presented at *Seventh European Workshop on Econometrics and Health Economics*, Helsinki.

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* **78**: 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.

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.

Jeong, J. and G.S. Maddala (1993), "A perspective on application of bootstrap methods in econometrics", In Maddala, G.S., C.R. Rao and H.D. Vinod, eds., *Handbook of Statistics, Vol. 11. (*Elsevier, Amsterdam) 519-543.

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

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.

López, A. (1998), "Unobserved heterogeneity and censoring in the demand for health care", *Health Economics* **7**: 429-437.

McClellan, M. and J.P. Newhouse (1997), "The marginal cost-effectiveness of medical technology: a panel instrumental variables approach", *Journal of Econometrics* **77**: 39-64.

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.

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.

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.

Mundlak, Y. (1978), "On the pooling of time series and cross section data", *Econometrica* **46**: 69-85.

Nanda, P. (1998), "The impact of women's participation in credit programs on the demand for quality health care in rural Bangladesh", presented at *Seventh European Workshop on Econometrics and Health Economics*, Helsinki.

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

Pudney, S. and M. Shields (1997), "Gender and racial inequality in pay and promotion in the internal labour market for NHS nurses", *Discussion Paper in Public Sector Economics no.97/4*, University of Leicester.

Robinson, P. (1988), "Root-N-consistent semiparametric regression", *Econometrica* **56**: 931-954.

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 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.

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

Vella, F. (1998), "Estimating models with sample selection bias", *Journal of Human Resources* **33**: 127-169.

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