Some Recent Developments in Microeconometrics

A. Colin Cameron
Department of Economics
University of California - Davis
accameron@ucdavis.edu

December 1, 2007

Abstract

This paper surveys methods that have been added to the microeconometrician's toolkit over the past twenty-five years, and some recent developments in these newer methods. These methods include GMM, empirical likelihood, simulation-based estimation, quantile regression, semiparametric estimation, robust inference, and bootstrap. The paper also considers estimation of marginal effects that can be given a causative interpretation, notably treatment effects, unobserved heterogeneity, and common data complications of sampling and missing and mismeasured data.

Keywords:

JEL Classification:

This paper is a draft of a paper to appear as a chapter in T.C. Mills and K. Patterson eds., Palgrave Handbook of Econometrics Volume 2: Applied Econometrics, forthcoming 2008. It draws considerably on Cameron and Trivedi (2005). Earlier versions were presented at the Japanese Statistical Society 75th Anniversary Symposium on Applied Microeconometrics, University of Tokyo, September, 2006, and at the 23rd Annual Summer Meeting of the Society for Political Methodology, U.C.-Davis, July, 2006. I have benefitted from comments of confrence participants, from sabbatical leave at U.C.-Berkeley, and from discussions with Bryan Graham, Michael Jansson, Jim Powell and Paul Ruud.

Contents

1 Introduction

Applied microeconometrics primarily applies regression methods to cross-section and longitudinal economics-related data.

Most often the goal is to obtain estimates of one or more marginal effects. A stereotypical example is estimation of the effect on earnings of a one-year increase in education. A simple approach is OLS estimation of a linear cross-section regression of log-earnings on years of schooling and other control variables. Potential complications include nonlinearity (with implications for estimation and statistical inference); endogeneity of the regressor schooling (that is chosen by the individual); unobserved individual heterogeneity (the marginal effect even after controlling for regressors may differ across individuals); and missing or mismeasured data.

In this paper I survey various methods to deal with these complications, most developed over the past twenty-five years. Some of these methods have already become well-established and command little current theoretical research. Other methods, especially those that are currently active areas of research, may or may not ultimately become part of the toolkit. An impetus for many of these methods is increased computing power and data availability.

The survey presumes the basic theory for least squares, maximum likelihood and instrumental variables estimation of nonlinear cross-section models and linear panel data models, as these were well established by the late 1970's. Section 2 presents a summary of identification. Newer estimation methods that enable use of richer models, notably generalized methods of moments, empirical likelihood, simulation-based methods (classical and Bayesian), quantile regression, and semiparametric estimation, are presented in Section 3. Recent developments in statistical inference, most notably robust standard errors and bootstrap methods, are presented in Section 4. Section 5 presents a range of methods that have been developed to obtain marginal effects that can be given a causative interpretation even when observational data are used. A fundamental change in thinking is the use of the potential outcomes framework and quasi-experimental approaches to tease out causation. Section 6 discusses methods to control for unobserved heterogeneity. Section 7 presents adjustments to standard methods that incorporate the practical data complications of survey sampling schemes, measurement error, and missing data.

The following notation is used. The typical observation is the i^{th} observation, with scalar dependent variable y_i, $k \times 1$ regressor vector $\mathbf{x}_i$, and, where relevant, $m \times 1$ instrument vector $\mathbf{z}_i$. Unless otherwise noted independence

over i is assumed. At times it is convenient to denote the i^{th} observation by $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ or $\mathbf{w}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i)$. The parameter vector in general is a $q \times 1$ vector $\boldsymbol{\theta}$. In some cases this is specialized to a $k \times 1$ parameter vector $\boldsymbol{\beta}$. Combining all N observations, $\mathbf{y}$ is the $N \times 1$ vector of dependent variables, and $\mathbf{X}$ is the $N \times K$ regressor matrix. The linear regression model is written as $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

The reader should be aware that this is a methods survey, rather than a literature survey. It is not possible to cite more than a few relevant references for each topic, leading to omission of the important contributions of many authors. More complete references are given in the relevant texts by Amemiya (1985), Greene (2003, first edition 1990), Davidson and MacKinnon (1993), Wooldridge (2002), and Cameron and Trivedi (2005). The most recent references given in this paper should provide a useful start to the current literature.

2 Identification

Introductory treatments of econometrics focus on specifying a parametric model for the conditional distribution $f(y|\mathbf{x}, \boldsymbol{\theta})$, or for the conditional mean, $\mathrm{E}[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta})$. Given specification of $f(\cdot)$ or $g(\cdot)$ and a sampling process such as random sampling or exogenous stratified sampling that provides no additional complication, the emphasis is on estimation of the parameters $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$, and on statistical inference based on these parameter estimates. Identification is discussed briefly in the context of rank conditions to ensure identification in linear simultaneous equations models. More generally, for nonlinear estimators in parameterized models Newey and McFadden (1994, p. 2134) state that "The identification condition for consistency of an extremum estimator is that the limit of the objective function has a unique maximum at the truth".

The literature on semiparametric modelling brings identification much more to the forefront. Identification asks the question whether a model, or key features of that model, can be estimated assuming an infinitely large sample is available and given the relevant sampling scheme. Only after identification is secured can one move on to estimation and inference given a finite sample. An example is a censored regression model, where we observe $y_i = y_i^*$ if $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ and $y_i = 0$ otherwise, and ask whether $\boldsymbol{\beta}$ is identified given assumptions on the distribution of u_i that fall short of complete parameterization of the distribution of u_i (such as assuming normality). In general there is no unified theory and identification conditions

vary with the model being considered and the sampling process. Also not all parameters may be identified. For example, regression coefficients may be identified only up to scale or an intercept may not be identified while slope parameters are. Pagan and Ullah (1999) provide many examples.

For nonparametric nonlinear simultaneous equations we consider model $\mathbf{r}(\mathbf{y}_i, \mathbf{x}_i) = \mathbf{u}_i$, where $\mathbf{y}$ and $\mathbf{u}$ are $G \times 1$ vectors and $\mathbf{x}$ is $K \times 1$. The model is nonparametric identified if it is possible to recover the unknown function $\mathbf{r}(\cdot)$ and the distribution of $\mathbf{u}$ from the joint distribution of $(\mathbf{y}, \mathbf{x})$. Matzkin (2005), building on Brown (1983), provides identification conditions when $\mathbf{u}$ is independent of $\mathbf{x}$.

Usually assumptions on the dgp are made sufficiently strong to ensure point identification or complete identification. But this is not always the case. Manski (1995, 2003, 2006) and related papers emphasize partial identification or set identification that merely provides bounds. For example, suppose data are observed with error but it is known that at most 25 percent of the data are mismeasured. Then the true population median, i.e. without measurement error, is bounded by the lower quartile (all the mismeasured observations are recorded as high values but are actually low values) and the upper quartile (all low values are actually high values). The attraction of partial identification compared to point identification is that it can rely on weaker assumptions about the dgp. The bounds can be wide, however, and additional information may permit tightening the bounds. Leading applications include Manski and Pepper (2000), Haile and Tamer (2003), and Blundell, Gosling, Ichimura, and Meghir (2007).

Finally it should be noted that while much of the literature focuses on identification of parameters, this may not be necessary. In particular, many studies in microeconometrics seek calculation of the marginal effect on the conditional mean of, say, the j^{th} regressor, $\partial E[y|\mathbf{x}]/\partial x_j|_{\mathbf{x}=\mathbf{x}^*}$. This can be achieved by nonparametric regression, such as the use of matching estimators in the treatment effects literature. And even where a model for $\mathrm{E}[y|\mathbf{x}]$ is posited, complete identification of $\mathrm{E}[y|\mathbf{x}]$ may not be necessary. For example, consider a linear panel fixed effects model where $\mathrm{E}[y_{it}|\mathbf{x}_{it}] = \mathbf{x}_{it}'\boldsymbol{\beta}$ and $\mathbf{x}_{it}$ includes a time-invariant variable, the k^{th} say, with $x_{ik} = x_k$. Then even if x_k is unobserved, fixed effects estimation provides consistent estimates of the remaining components of $\boldsymbol{\beta}$ and hence the marginal effect. The marginal effects of other conditional moments can also be of interest. Examples include the conditional median, conditional quantiles and conditional variance.

3 Estimation

Generalized method of moments, which provides a quite general framework for estimation, is presented in Section 3.1. Empirical likelihood, an adaptation of GMM with different finite sample properties is presented in Section 3.2. Simulation methods that permit classical and Bayesian methods to be applied to a much wider range of models are presented in, respectively, Sections 3.3 and 3.4. Quantile regression and semiparametric methods are presented in, respectively, Sections 3.5 and 3.6.

3.1 Generalized Method of Moments

The starting point for GMM is the moment condition

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \tag{1}$$

where $\mathbf{h}(\cdot)$ is an $r \times 1$ vector.

The analogy principle, emphasized by Manski (1988) who attributes it to Goldberger, proposes estimation using the sample analog of the population condition (1). In the just-identified case this leads to the method of moments (MM) estimator $\widehat{\boldsymbol{\theta}}_{\text{MM}}$ that solves $N^{-1}\sum \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}$.

In the over-identified case that $N^{-1}\sum_i \partial \mathbf{h}_i/\partial \boldsymbol{\theta}'$ has rank greater than q, there are more moment conditions than parameters. Then Hansen (1982) proposed the generalized method of moments estimator $\widehat{\boldsymbol{\theta}}_{\text{GMM}}$ that minimizes the quadratic form

$$Q(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum\nolimits_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right]' \mathbf{W}_N \left[\frac{1}{N}\sum\nolimits_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right], \tag{2}$$

where $\mathbf{W}_N$ is an $r \times r$ symmetric full rank weighting matrix that is usually data dependent. Under appropriate assumptions, including that (1) holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the GMM estimator $\widehat{\boldsymbol{\theta}}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance matrix of sandwich form

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}_{\text{GMM}}] = \frac{1}{N}\left(\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{G}}\right)^{-1}\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{S}}\mathbf{W}_N\widehat{\mathbf{G}}\left(\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{G}}\right)^{-1}, \tag{3}$$

where $\widehat{\mathbf{G}} = N^{-1}\sum_i \partial \mathbf{h}_i/\partial \boldsymbol{\theta}'\big|_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{S}}$ is a consistent estimate of $\mathbf{S}_0 = \text{plim}\,\frac{1}{N}\sum_i\sum_j \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)\mathbf{h}(\mathbf{w}_j, \boldsymbol{\theta}_0)'$. Given independence over i, $\widehat{\mathbf{S}}$ simplifies to $\widehat{\mathbf{S}} = \frac{1}{N}\sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$, while for clustered observations adaptations similar to those given in Section 4.1 are used.

A leading example is instrumental variables estimation. The condition that instruments $\mathbf{z}_i$ are uncorrelated with the error term $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ in a linear regression model implies that $\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. In the just-identified case the MM estimator solves $\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$, which yields the instrumental variables estimator. In the over-identified case the GMM estimator minimizes $[\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})]' \mathbf{W}_N [\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. The two-stage least squares estimator is the special case $\mathbf{W}_N = \left[N^{-1}\sum_i \mathbf{z}_i\mathbf{z}_i'\right]^{-1}$.

For just-identified models the GMM estimator reduces to the MM estimator regardless of the choice of $\mathbf{W}_N$. For over-identified models the most efficient GMM estimator based on the moment conditions (1), called the optimum GMM (OGMM) or two-step GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$, sets $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$ where $\widehat{\mathbf{S}}$ is a consistent estimate of $\mathbf{S}_0$. Then the OGMM estimator has estimated asymptotic variance

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}_{\mathrm{OGMM}}] = N^{-1}(\widehat{\mathbf{G}}'\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{G}})^{-1}.$$

In practice, however, it is found that the optimal GMM estimator suffers from small sample bias, see Altonji and Segal (1996), and other simpler choices of $\mathbf{W}_N$ may be better. This has spawned an active literature, see Windmeijer (2005), including that on empirical likelihood given in Section 3.2.

There are several attractions to GMM. First, it provides a natural extension of instrumental variables methods in over-identified models from linear to nonlinear models, and can be viewed as a generalization of nonlinear 2SLS. Second, it provides a unifying framework to estimation as it nests many estimation procedures, including LS, with $\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, and ML, with $\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, as special cases. Third, it views estimation as a sample analog to population moment conditions, the analogy principle emphasized by Manski (1988). Fourth, taking this view leads naturally to conditional moment tests, see Section 4.2, that lead to model moment specification tests based on model moment conditions that are not exploited in estimation.

A method closely related to GMM though less used is minimum distance estimation. Suppose that the relationship between q structural parameters and $r > q$ reduced form parameters is that $\boldsymbol{\pi} = \mathbf{g}(\boldsymbol{\theta})$. And suppose that we have a consistent estimate $\widehat{\boldsymbol{\pi}}$ of the reduced form parameters. An obvious estimator is $\widehat{\boldsymbol{\theta}}$ such that $\widehat{\boldsymbol{\pi}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$, but this is infeasible since $q < r$. Instead the minimum distance estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MD}}$ minimizes with respect to $\boldsymbol{\theta}$ the objective function

$$Q_N(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta}))'\mathbf{W}_N(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta})), \tag{4}$$

where $\mathbf{W}_N$ is an $r \times r$ weighting matrix. The optimal MD estimator uses weighting matrix $\mathbf{W}_N = \widehat{\mathrm{V}}[\widehat{\boldsymbol{\pi}}]^{-1}$ in (4). This estimator is used mainly in panel data analysis, see Chamberlain (1982, 1984), especially in estimation of covariance structures, see Abowd and Card (1989).

The statistics literature rarely uses the GMM framework. This may be because GMM is particularly useful for overidentified models, notably IV with surplus instruments, that are much more often used in econometrics. Instead, for nonlinear models the statistics literature emphasizes the more restrictive generalized linear models and generalized estimating equations frameworks, see McCullagh and Nelder (1983, 1989).

3.2 Empirical Likelihood

Empirical likelihood is based on the same moment conditions as GMM, but is a different estimation method with second-order asymptotic properties that differ from GMM so that the estimator may have better finite sample properties.

Let $\pi_i = f(y_i|\mathbf{x}_i)$ denote the probability that the i^{th} observation on y takes the realized value y_i. The empirical likelihood (EL) approach, introduced by Owen (1988), maximizes the empirical log-likelihood function

$$Q_N(\pi_1, ..., \pi_N) = N^{-1} \sum_i \ln \pi_i, \tag{5}$$

subject to any model constraints.

With no model the only constraint is that probabilities sum to one. This leads to maximum EL estimates $\widehat{\pi}_i = 1/N$, so the estimated density function $\widehat{f}(y|\mathbf{x})$ has mass $1/N$ at each of the realized values y_i, $i = 1, ..., N$, and the resulting distribution function estimate is just the usual empirical distribution function.

With a model introduced attention focuses on the estimates for parameters of that model. In the simplest case of estimation of a common population mean μ, the maximum EL estimate can be shown to be the sample mean. A more general example is to specify a model that imposes r moment conditions

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \tag{6}$$

the same condition as in (1) for MM or GMM estimation. The empirical likelihood approach maximizes the empirical likelihood function $N^{-1} \sum_i \ln \pi_i$ subject to the constraint $\sum_i \pi_i = 1$, since probabilities sum to one, and the additional sample constrained based on the population moment condition

(6) that

$$\sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{7}$$

Thus we maximize with respect to $\boldsymbol{\pi} = [\pi_1 ... \pi_N]'$, η, $\boldsymbol{\lambda}$, and $\boldsymbol{\theta}$ the Lagrangian

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ln \pi_i - \eta \left(\sum_{i=1}^{N} \pi_i - 1 \right) - \boldsymbol{\lambda}' \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}), \tag{8}$$

where the Lagrangian multipliers are a scalar η and an $r \times 1$ column vector $\boldsymbol{\lambda}$.

This maximization is not straightforward. First concentrate out the N parameters $\pi_1, ..., \pi_N$. Differentiating $\mathcal{L}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta})$ with respect to π_i yields $1/(N\pi_i) - \eta - \boldsymbol{\lambda}'\mathbf{h}_i = 0$. Then find $\eta = 1$ by multiplying by π_i and summing over i and using $\sum_i \pi_i \mathbf{h}_i = \mathbf{0}$. It follows that the N Lagrangian multipliers $\pi_i(\boldsymbol{\theta}, \boldsymbol{\lambda}) = 1/[N(1 + \boldsymbol{\lambda}'\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))]$. The problem is now reduced to a maximization problem with respect to $(r + q)$ variables $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the Lagrangian multipliers associated with the r moment conditions (7) and the q parameters $\boldsymbol{\theta}$. Solution at this stage requires numerical methods, even for just-identified models with $r = q$. After some algebra, the log-likelihood function evaluated at $\boldsymbol{\theta}$ is

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\theta}) = -N^{-1} \sum_{i=1}^{N} \ln[N(1 + \boldsymbol{\lambda}(\boldsymbol{\theta})'\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))]. \tag{9}$$

The maximum empirical likelihood (MEL) estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MEL}}$ maximizes this function with respect to $\boldsymbol{\theta}$.

Qin and Lawless (1994) show that the MEL estimator has the same limit distribution as the optimal GMM estimator. In finite samples, however, $\widehat{\boldsymbol{\theta}}_{\mathrm{MEL}}$ differs from $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$. Furthermore, inference can be based on sample estimates $\widehat{\mathbf{G}} = \sum_i \widehat{\pi}_i \partial \mathbf{h}_i / \partial \boldsymbol{\theta}' \big|_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{S}} = \sum_i \widehat{\pi}_i \mathbf{h}_i(\widehat{\boldsymbol{\theta}}) \mathbf{h}_i(\widehat{\boldsymbol{\theta}})'$ that weight by the estimated probabilities $\widehat{\pi}_i$ rather than the proportions $1/N$.

Imbens (2002) and Kitamura (2006) provide recent surveys of empirical likelihood. Objective functions other than $N^{-1} \sum_i \ln \pi_i$ may be used, such as $N^{-1} \sum_i \pi_i \ln \pi_i$. Newey and Smith (2004) show that MEL has better second-order asymptotic properties than GMM, and it appears that using the weights $\widehat{\pi}_i$ in forming $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{S}}$ leads to improved finite sample performance. These results suggest that MEL may be better than optimal GMM in applications, but MEL is not yet widely used as it is viewed to be computationally more burdensome; see Imbens (2002) for a discussion.

3.3 Simulation-Based ML and MM Estimation

ML estimation requires specification of a density. In some cases the density includes an integral for which a closed-form solution does not exist, so that conventional ML is not possible. Simulation-based estimation methods enable ML estimation in this case by approximating the integral by Monte Carlo integration, making many draws from an appropriate distribution.

Specifically, we suppose that the conditional density of y given regressors $\mathbf{x}$ and parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ is an integral

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \int f(y|\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}_1) g(\mathbf{u}|\boldsymbol{\theta}_2) d\mathbf{u}, \tag{10}$$

where $f(y|\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}_1)$ which depends in part on unobservables $\mathbf{u}$ is of closed form, but there is no closed form for the desired density $f(y|\mathbf{x}, \boldsymbol{\theta})$.

A leading example is unobserved heterogeneity. Then $\boldsymbol{\theta}_1$ denotes parameters of intrinsic interest, $\mathbf{u}$ denotes unobserved heterogeneity which may depend on unknown parameters $\boldsymbol{\theta}_2$, and the integral will not have a closed form solution except in some special cases. A second example is the multinomial probit model. Then $\boldsymbol{\theta}_1$ denotes regression parameters, $\mathbf{u}$ denotes error term in a latent model that may have unknown error variances and covariances $\boldsymbol{\theta}_2$, and, given m alternatives, the probability that a specific alternative is chosen is given by an $(m-1)$-dimensional integral that has no closed-form solution.

If the integral is of low dimension, then numerical integration by Gaussian quadrature may provide a reasonable approximation to $f(y|\mathbf{x}, \boldsymbol{\theta})$. But these methods can work poorly in higher dimensions often encountered in practice. For example, for multinomial probit numerical methods are felt to work poorly if there are more than four alternatives.

Instead, the maximum simulated likelihood (MSL) method makes many draws of the unobservables $\mathbf{u}$ from density $g(\mathbf{u}|\boldsymbol{\theta}_2)$. The MSL estimator maximizes the simulated log-likelihood function

$$\widehat{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta}), \tag{11}$$

where $\widehat{f}(\cdot)$ is the Monte Carlo estimate or simulator

$$\widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} \widetilde{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^{s}, \boldsymbol{\theta}), \tag{12}$$

where $\mathbf{u}_i^{(S)} = (\mathbf{u}_i^1, ..., \mathbf{u}_i^S)$ denotes S draws with marginal density $g(\mathbf{u}_i|\boldsymbol{\theta}_2)$, and $\widetilde{f}(\cdot)$ is a subsimulator such as $f(y|\mathbf{x}, \mathbf{u}^s, \boldsymbol{\theta}_1)$. Many possible simulators may be used - the essential requirement is that $\widehat{f}_i \xrightarrow{p} f_i$ as $S \to \infty$. The MSL estimator is consistent and asymptotically equivalent to the ML estimator, provided that $S \to \infty$, in addition to the usual assumption that $N \to \infty$, with $\sqrt{N}/S \to \infty$ so that S grows at rate slower than N.

The MSL estimator opens up the possibility of using a much wider range of parametric models, such as richer models for unobserved heterogeneity that may be more robust to model misspecification. At the same time the method can be computationally demanding. An early application of MSL was by Lerman and Manski (1981), for the multinomial probit model. Then $I \times N \times S$ draws of $\mathbf{u}_i^s$ are made if analytical derivatives are used, where I is the number of iterations, and even more draws are needed if numerical derivatives are used.

One common application of MSL is to models with unobserved heterogeneity, where implicitly we have treated the heterogeneity as being continuously distributed. An alternative is to treat heterogeneity as being discretely distributed, often with just two or three points of support. Such finite mixture or latent class models are especially popular in the duration and count (number of health services) literatures; see Meyer (1990) and Deb and Trivedi (2002). These models can be more easily estimated using quasi-Newton methods or the expectation maximization algorithm.

The MSL can be extended to method of moments and generalized method of moments estimation. In that case theory leads to a moment condition $\mathrm{E}[m(y_i|\mathbf{x}_i, \boldsymbol{\theta})] = 0$, where $m(\cdot)$ is a scalar for simplicity, but there is no closed form expression for $m(y, \mathbf{x}, \boldsymbol{\theta})$. Instead $m(y, \mathbf{x}, \boldsymbol{\theta})$ is an integral

$$m(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int h(y_i|\mathbf{x}_i, \mathbf{u}_i, \boldsymbol{\theta}_1) g(\mathbf{u}_i|\boldsymbol{\theta}_2) d\mathbf{u}_i, \tag{13}$$

for some functions $h(\cdot)$ and $g(\cdot)$, where $m(\cdot)$ has no closed form. Let $\widehat{m}_i = \widehat{m}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta})$ be an simulator for $m(y_i, \mathbf{x}_i, \boldsymbol{\theta})$.Then the method of simulated moments (MSM) estimator uses $\widehat{m}_i$ in place of m_i in GMM estimation. A key result, due to McFadden (1989) and Pakes and Pollard (1989), is that the MSM estimator is consistent for $\boldsymbol{\theta}$ as $N \to \infty$ even if S is very small, provided that an unbiased simulator is used, meaning $\mathrm{E}[\widehat{m}_i] = m_i$. Furthermore small S may lead to little loss of precision. In the special case that $\widehat{m}(\cdot)$ is the frequency simulator, the MSM estimator has variance $(1 + (1/S))$ times that of the MM estimator. The biggest loss in efficiency is that compared to the MSL estimator which requires $S \to \infty$ as

an unbiased simulator for the density does not lead to an unbiased simulator for the log density and its derivative.

There are several subtleties in use of MSL and related estimators. Book references are Gourieroux and Monfort (1996), who also discuss indirect inference, and Train (2003) who focuses on applications to multinomial choice. First, because the simulated likelihood is usually maximized by iterative gradient methods, the simulator $\widehat{f}_i$ should be differentiable (or smooth) in $\boldsymbol{\theta}$. For example, for limited dependent variables models with normal errors the GHK simulator is often used. Second, to enable convergence and avoid "chatter" the same underlying random numbers used to obtain $\mathbf{u}_i^S$ should be used at each iteration. Third, the draws from $g(\mathbf{u}_i|\boldsymbol{\theta}_2)$ need not be independent. For example, better approximation for given S may be obtained by using dependent quasi-random numbers, such as Halton sequences, rather than independent pseudo-random numbers, and by use of antithetic sampling. Fourth it may be difficult to make draws from $\mathbf{u}_i^S$ using standard methods such as inverse transformation and accept-reject methods. Then newer Markov chain Monte Carlo methods, widely used in Bayesian analysis, may be used.

3.4 Simulation-Based Bayesian Analysis

Let $L(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$ denote the sample joint density or likelihood, and $\pi(\boldsymbol{\theta})$ denote the prior distribution. Then the posterior density for $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{L(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{X})}, \tag{14}$$

where $f(\mathbf{y}|\mathbf{X}) = \int_{R(\boldsymbol{\theta})} L(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $R(\boldsymbol{\theta})$ denotes the support of $\pi(\boldsymbol{\theta})$. Because the denominator $f(\mathbf{y}|\mathbf{X})$ is free of $\boldsymbol{\theta}$, we can more simply write

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{15}$$

where for notational simplicity we suppress the regressors $\mathbf{X}$. The posterior is proportional to the product of the likelihood and prior.

The heart of Bayesian analysis is the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. In the simplest cases a closed form expression for this exists. For example, if $\mathbf{y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and known variance and the prior for $\boldsymbol{\beta}$ is the normal with specified mean and variance, then the posterior is normal.

But for most models, especially standard nonlinear regression models, the posterior is unknown. One approach is to then obtain key moments, such as the posterior mean $\mathrm{E}[\boldsymbol{\theta}] = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ using Monte Carlo integration

methods that do not require draws from $p(\boldsymbol{\theta}|\mathbf{y})$. In particular importance sampling methods can be used; see Kloek and van Dijk (1978) and Geweke (1989).

The more modern approach is to instead to obtain many draws, say $\widehat{\boldsymbol{\theta}}^1, ..., \widehat{\boldsymbol{\theta}}^S$ from $p(\boldsymbol{\theta}|\mathbf{y})$. Then the posterior mean can then be estimated by $S^{-1}\sum_{s=1}^{S}\widehat{\boldsymbol{\theta}}^s$, and other quantities of interest, such as the distribution of marginal effects in a nonlinear model, can be similarly computed. The key ingredient is the recent development of methods to obtain draws of $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ even when $p(\boldsymbol{\theta}|\mathbf{y})$ is unknown, see Gelfland and Smith (1990).

The starting point is the Gibbs sampler. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ and suppose that it is possible to draw from the conditional posteriors $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$, even though it is not possible to draw from $p(\boldsymbol{\theta}|\mathbf{y})$. The Gibbs sampler obtains draws from $p(\boldsymbol{\theta}|\mathbf{y})$ by making alternating draws from each conditional distribution. Thus given an initial value $\boldsymbol{\theta}_2^{(0)}$, we obtain $\boldsymbol{\theta}_1^{(1)}$ by drawing from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(0)}, \mathbf{y})$, then $\boldsymbol{\theta}_2^{(1)}$ by drawing from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(1)}, \mathbf{y})$, then $\boldsymbol{\theta}_1^{(2)}$ by drawing from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(1)}, \mathbf{y})$, and so on. When repeated many times it can be shown that this process ultimately leads to draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y})$, even though in general $p(\boldsymbol{\theta}|\mathbf{y}) \neq p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}) \times p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$. The sampler is an example of a Markov chain Monte Carlo method. The term Markov chain is used because the procedure sets up a Markov chain for $\boldsymbol{\theta}$ whose stationary distribution can be shown to be the desired posterior $p(\boldsymbol{\theta}|\mathbf{y})$. The method extends immediately to more partitions for $\boldsymbol{\theta}$. For example, if $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2' \ \boldsymbol{\theta}_3']'$ then we need to be able to make draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \mathbf{y})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \mathbf{y})$, and $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{y})$.

In many applications some of the conditional posteriors are unknown, in which case MCMC methods other than the Gibbs sampler need to be used. A standard method is the Metropolis-Hastings algorithm which uses a trial or jumping distribution. The Gibbs sampler can be shown to be an example of a Metropolis-Hastings algorithm, one with relatively fast convergence.

The MCMC methods in principle permit Bayesian analysis to be applied to a very wide range of models. In practice, there is an art to ensuring that the chain converges in a reasonable amount of computational time. The first B draws of $\boldsymbol{\theta}$ are discarded, where B is chosen to be large enough that the Markov chain has converged. The remaining S draws of $\boldsymbol{\theta}$ are then used. Various diagnostic methods exist to indicate convergence though these do not guarantee convergence. MCMC methods yield correlated draws from $p(\boldsymbol{\theta}|\mathbf{y})$, rather than independent draws, but this correlation only effects the precision of posterior analysis and often the correlation is low. Many

Bayesian models include both components with closed from solutions for the posterior and components that require use of MCMC methods – the Gibbs sampler, if possible, and failing that the MH algorithm with hopefully good choice of jumping distribution.

Bayesian methods are particularly attractive in models entailing latent variables, such as Tobit models, see Chib (1992, 2001), and multinomial probit models, see McCulloch, Polson, and Rossi (2000). Then data augmentation, see Tanner and Wong (1987), is used. A recent application is that by Geweke, Gowrisankaran, and Town (2003). Recent econometrics books are Koop (2003), Lancaster (2004), and Koop, Poirier, and Tobias (2007).

Bayesian inference is quite different from frequentist inference, and the difference has provoked strong philosophical debates. Frequentists when confronted by numerically challenging likelihood functions can obtain ML estimates of $\boldsymbol{\theta}$ by using the preceding Bayesian MCMC methods with a diffuse prior for $\boldsymbol{\theta}$ specified, in which case the sample information dominates, and then proceed to use frequentist inferential methods. Whether the new Bayesian numerical methods will lead to greater use of Bayesian inferential methods is an open question.

3.5 Quantile Regression

In the iid case quantiles, such as deciles and quartiles, are often used to summarize the distribution of income, earnings and wealth. Quantile regression is an extension to the regression case where, for example interest may lie in different response of earnings to education at different levels of earnings.

The least squares estimator maximizes the sum of squared residuals, but alternative functions of the residuals can be considered. In particular, the least absolute deviations (LAD) estimator minimizes the sum of absolute residuals $\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$. In the iid case, with $\mathbf{x}_i'\boldsymbol{\beta} = \beta$, the resulting estimate of β is the sample median.

More generally we can consider estimation of quantiles other than the median. The q^{th} quantile regression estimator $\widehat{\boldsymbol{\beta}}_q$ minimizes over $\boldsymbol{\beta}_q$

$$Q_N(\boldsymbol{\beta}_q) = \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}}^{N} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}}^{N} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|,$$

where we use $\boldsymbol{\beta}_q$ rather than $\boldsymbol{\beta}$ to make clear that different choices of q estimate different values of $\boldsymbol{\beta}$. The special case $q = 0.5$ is the LAD estimator. The objective function is not differentiable, so linear programming

methods are used rather than more familiar gradient methods. These enable relatively fast computation of $\widehat{\boldsymbol{\beta}}_q$. The quantile regression estimator is consistent and asymptotically normal, but estimation of the variance of $\widehat{\boldsymbol{\beta}}_q$ requires estimation of $f_{u_q}(0|\mathbf{x})$, the conditional density of the error term $u_q = y - \mathbf{x}'\boldsymbol{\beta}_q$ evaluated at $u_q = 0$. An easier method is to instead obtain bootstrap standard errors for $\widehat{\boldsymbol{\beta}}_q$ using a paired bootstrap.

Quantile regression was proposed by Koenker and Bassett (1978). Powell (1984, 1986) adapted the method to permit consistent estimation in censored linear regression models without specification of the distribution of the errors. Buchinsky (1994) provided a much-cited application. Chernozhukov and Hansen (2005) propose a LAD IV estimator. Angrist, Chernozhukov and Ferandez-Val (2006) provide interpretation of the quantile regression when the quantile function is misspecified. Koenker and Hallock (2001) and Koenker (2005) provide summaries of the quantile literature.

3.6 Nonparametric and Semiparametric Methods

Consider the regression model

$$\mathrm{E}[y_i|\mathbf{x}_i] = m(\mathbf{x}_i), \tag{16}$$

where the function $m(\mathbf{x})$ is unspecified. Nonparametric regression provides a consistent estimate of $m(\mathbf{x})$. At the specific point $\mathbf{x} = \mathbf{x}_0$, $m(\mathbf{x}_0)$ can be estimated by taking a local weighted average of y_i over those observations with $\mathbf{x}_i$ in a neighborhood of $\mathbf{x}_0$. There are many variations on this approach, including kernel regression, nearest neighbors regression, local linear, local polynomial, Lowess, smoothing spline and series estimators. Because a local average is taken less than N observations are effectively used at any point $\mathbf{x}_0$, so $\widehat{m}(\mathbf{x}_0) \xrightarrow{p} m(\mathbf{x}_0)$ at rate less than the usual $N^{-1/2}$.

Fully nonparametric regression works best in practice when there is just a single regressor. Even then, empirical results vary greatly with the choice of bandwidth or window width that defines the size of the neighborhood. Unlike kernel density estimation, "plug-in" estimates of the bandwidth work very poorly. Cross-validation is commonly-used to select the bandwidth, but this method is by no means perfect.

There is no theoretical obstacle to using nonparametric regression when there are many regressors. But in practice nonparametric methods usually work poorly with more than very few regressors. The optimal convergence rate using mean-squared error as a criterion is $N^{-2/(\dim[\mathbf{x}]+4)}$. As $\dim[\mathbf{x}]$ increases the convergence rate deceases, a curse of dimensionality that arises

because the local averages will be made over fewer observations. For example, if we averaged over 10 bins with one regressor we may wish to average over $10^2 = 100$ bins when there are two regressors. This problem is less severe when some regressors take only a few values, such as binary indicator variables. Racine and Li (2004) present results for kernel regression when some regressors are discrete and some are continuous.

The microeconometrics literature focuses on semiparametric methods that overcome the curse of dimensionality by partially parameterizing a model, so that there is a mix of parametric and nonparametric components. The maximum score estimator for the binary choice model of Manski (1975) is a very early example. A first step is to determine whether a model is identified. Ideally $\sqrt{N}$-consistent and asymptotically normal estimates of the parameters can be obtained. These should be fully efficient in that they attain semi-parametric efficiency bounds, see Chamberlain (1987), Newey (1990), and Severini and Tripathi (2001), that are extensions of Cramer-Rao lower bounds or the Gauss-Markov theorem.

There are many semiparametric models, and for each model there can be several different ways to obtain estimators. We present two commonly-used semiparametric models in econometrics that are also the building blocks towards more general models.

The partial linear model specifies the conditional mean to be the usual linear regression function plus an unspecified nonlinear component, so

$$\mathrm{E}[y_i|\mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \lambda(\mathbf{z}_i), \tag{17}$$

where the scalar function $\lambda(\cdot)$ is unspecified. An example is the estimation of a demand function for electricity, where $\mathbf{z}$ reflects time-of-day or weather indicators such as temperature. A second example is a sample selection model where $\lambda(\mathbf{z})$ is the expected value of a model error, conditional on the sample selection rule. In applications interest may lie in $\boldsymbol{\beta}$, $\lambda(\mathbf{z})$ or both.

Various estimators for the partial linear model have been proposed. The differencing method proposed by Robinson (1988) estimates $\boldsymbol{\beta}$ by OLS regression of $(y_i - \widehat{m}_{yi})$ on $(\mathbf{x}_i - \widehat{\mathbf{m}}_{\mathbf{x}i})$, where $\widehat{m}_{yi}$ and $\widehat{\mathbf{m}}_{\mathbf{x}i}$ are predictions from nonparametric regression of, respectively, y and $\mathbf{x}$ on $\mathbf{z}$. Robinson used kernel estimates that may need to be oversmoothed. Other methods that additionally estimate $\lambda(z)$, at least for scalar z, include a generalization of the cubic smoothing spline estimator, and using a series approximation for $\lambda(z)$.

The single-index model specifies the conditional mean to be an unknown

scalar function of a linear combination of the regressors, with

$$\mathrm{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i'\boldsymbol{\beta}), \tag{18}$$

where the scalar function $g(\cdot)$ is unspecified and the parameters $\boldsymbol{\beta}$ are then only identified up to location and scale. An example is a binary choice model with $\Pr[y = 1|\mathbf{x}] = g(\mathbf{x}'\boldsymbol{\beta})$ where $g(\cdot)$ is unknown. The single-index formulation is attractive as the marginal effect of a change in the jth regressor is $g'(\mathbf{x}_i'\boldsymbol{\beta})\beta_j$, so that the ratio of parameter estimates equals the ratio of marginal effects.

Estimators for the single-index model include an average derivative estimator, a density weighted average derivative estimator, see Powell, Stock and Stoker (1989), and semiparametric least squares.

Microeconometricians have focused on semiparametric estimation for limited dependent variable models - binary choice with unspecified function for the probabilities, censored regression and sample selection. The literature is vast. References include the book by Pagan and Ullah (1999) and the applied study by Bellemare, Melenberg and Van Soest (2002). Nonparametric and semiparametric methods are also used in the treatment effects literature detailed in Section 5.1.

4 Statistical Inference

Robust statistical inference, presented in Section 4.1, presents Wald tests based on robust standard errors that rely on distributional assumptions that are as weak as possible. Other developments in hypothesis testing and model specification testing are presented in Section 4.2. These methods rely on asymptotic results that provide only an approximation in typical finite sample sizes. The bootstrap, detailed in Section 4.3, provides an alternative way to compute asymptotic approximations. Furthermore it can in some cases additionally provide a more accurate asymptotic approximation.

4.1 Robust Inference for Wald Tests

We begin with the cross-section case of independent observations, before moving to clustered observations which includes short panels.

Consider an m-estimator $\widehat{\boldsymbol{\theta}}$ that maximizes with respect to $\boldsymbol{\theta}$ the objective function $Q_N(\boldsymbol{\theta}) = N^{-1}\sum_i q(y_i, \mathbf{x}_i, \boldsymbol{\theta})$. For maximum likelihood estimation $q(\cdot)$ is the log-density, and for least squares estimation $q(\cdot)$ is minus the

squared error (or a rescaling of this). The m-estimator solves the first-order conditions

$$N^{-1}\sum\nolimits_i \mathbf{h}_i(\boldsymbol{\theta}) = \mathbf{0}, \tag{19}$$

where $\mathbf{h}_i(\boldsymbol{\theta}) = \partial q(y_i, \mathbf{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$**.** Under suitable assumptions, notably that $\mathrm{E}[\mathbf{h}_i(\boldsymbol{\theta})] = 0$ in the population, it can be shown that $\widehat{\boldsymbol{\theta}}$ is $\sqrt{N}-$consistent, with limit distribution

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{\prime -1}], \tag{20}$$

where $\mathbf{A}_0 = \mathrm{plim} N^{-1}\sum_i \partial \mathbf{h}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'\big|_{\boldsymbol{\theta}_0}$, and $\mathbf{B}_0 = \mathrm{plim} N^{-1}\sum_i \mathbf{h}_i(\boldsymbol{\theta}_0)\mathbf{h}_i(\boldsymbol{\theta}_0)'$, and $\boldsymbol{\theta}_0$ is the value of $\boldsymbol{\theta}$ in the data generating process (dgp).

In practice we base inference on $\widehat{\boldsymbol{\theta}}$ being asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance matrix of sandwich form

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = N^{-1}\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{\prime -1}, \tag{21}$$

where $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are consistent estimates of $\mathbf{A}_0$ and $\mathbf{B}_0$. The Wald test statistic for $H_0 : \theta_j = r$ is then $W = (\widehat{\theta}_j - r)/s_j$ where s_j is the j^{th} diagonal entry of $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$ and $W \overset{a}{\sim} \mathcal{N}[01]$ under H_0. More generally to test $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ we use $W = \mathbf{h}(\widehat{\boldsymbol{\theta}})'(\widehat{\mathbf{R}}'\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]\widehat{\mathbf{R}})^{-1}\mathbf{h}(\widehat{\boldsymbol{\theta}})$ where $\widehat{\mathbf{R}} = \partial \mathbf{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\widehat{\boldsymbol{\theta}}}$ and $W \overset{a}{\sim} \chi^2(rank[\mathbf{R}])$ under H_0.

There are several possible ways to form $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, depending in part on the strength of distributional assumptions made. Robust variance estimates are those that rely on minimal distributional assumptions, provided $N \to \infty$.

Given data independent over i, the robust variance matrix estimate uses

$$\begin{aligned} \widehat{\mathbf{A}} &= N^{-1}\sum\nolimits_i \frac{\partial \mathbf{h}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\widehat{\boldsymbol{\theta}}}, \\ \widehat{\mathbf{B}} &= N^{-1}\sum\nolimits_i \mathbf{h}_i(\widehat{\boldsymbol{\theta}})\mathbf{h}_i(\widehat{\boldsymbol{\theta}})'. \end{aligned} \tag{22}$$

The resulting standard errors are called robust standard errors. In some cases the Hessian $\widehat{\mathbf{A}}$ in (22) may be replaced by the expected Hessian, and $\widehat{\mathbf{B}}$ may use a degrees-of- freedom correction such as $(N-q)^{-1}$ rather than N^{-1}.

A leading example is the heteroskedastic-consistent estimate of the variance-covariance matrix of the ordinary least squares (OLS) estimator. Then $q_i(\boldsymbol{\beta}) = -\frac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$, where the multiple $\frac{1}{2}$ is added for convenience, so that $\mathbf{h}_i(\boldsymbol{\beta}) = \partial q_i/\partial \boldsymbol{\beta} = (y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i$, and $\partial \mathbf{h}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}' = -\mathbf{x}_i\mathbf{x}_i'$. It follows that

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}] = \left[\sum\nolimits_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum\nolimits_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right]\left[\sum\nolimits_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1}, \tag{23}$$

where $\widehat{u}_i = (y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})$.

For ML estimation use of (22) relaxes the traditional information matrix equality assumption that $\mathbf{A}_0 = -\mathbf{B}_0$, which gives the simplification $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} = -\mathbf{A}_0^{-1}$. Failure of the information matrix equality, however, will generally imply inconsistency of the MLE. (A notable exception is models such as Poisson and logit and probit with specified density in the linear exponential family and correctly specified conditional mean function). In that case $\boldsymbol{\theta}_0$ needs to be reinterpreted as a "pseudo-true value", which is the value of $\boldsymbol{\theta}$ that maximizes the probability limit of $1/N$ times the log-likelihood function.

The estimates in (22) can be extended to clustered data. In that case observations are grouped into clusters, with correlation permitted within cluster but independence assumed across clusters. An example is panel data where the cluster unit is the individual, observations for a given individual over time are correlated, but observations across individuals are independent. Failure to control for clustering can lead to greatly under-estimated standard errors. Let $c = 1, ..., C$ denote clusters and let $j = 1, ..., N_c$ denote the N_c observations in cluster c. Then the cluster-robust variance matrix estimate is (21) where $\widehat{\mathbf{A}}$ is again given in (22) but now

$$\widehat{\mathbf{B}} = N^{-1}\sum_{c=1}^{C}\sum_{j=1}^{N_c}\sum_{k=1}^{N_c}\mathbf{h}_{jc}(\widehat{\boldsymbol{\theta}})\mathbf{h}_{kc}(\widehat{\boldsymbol{\theta}})'. \tag{24}$$

This estimator permits both error heteroskedasticity and quite flexible error correlation within cluster. It has largely supplanted use of a more restrictive random effects or error components model, though it does require $C \to \infty$.

The theory for robust inference is well-established and, in the independent observations case at least, is well incorporated into microeconometrics practice. In particular, for LS problems it is standard to estimate by OLS and then use robust standard errors, even though there may be efficiency loss compared to doing feasible GLS. Note, however, that one can still employ feasible GLS but then compute robust standard errors that guard against misspecification of the model for the error variance matrix.

For independent errors the key early reference is White (1980) who proposed the special case (23). Robust standard errors have been applied to many estimators, including instrumental variables and generalized method of moments (see (3) and Newey and West, 1987a). Amemiya (1985) and Newey and McFadden (1994) provide quite general treatments of inference and estimation; see also White (1984, 200?). For clustered errors various references are given by Cameron, Gelbach and Miller (2006a, 2006b) who consider, respectively, finite-sample corrections when there are few clusters

and extensions to multi-way clustering.

4.2 Hypothesis Tests and Model Specification Tests

For hypotheses on parameters of the form

$$\begin{aligned} H_0 &: \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0} \\ H_a &: \mathbf{h}(\boldsymbol{\theta}) \neq \mathbf{0}, \end{aligned}$$

the classical tests in the likelihood framework are Wald, Lagrange multiplier (or score) tests, and the likelihood ratio test. For correctly specified likelihood function these tests are first-order asymptotically equivalent under the null hypothesis and under local alternatives, so choice between them is one of convenience.

More recent work has focused on finite-sample properties of the tests and generalization to the non-likelihood framework.

The Wald test has become the most popular of these three tests, as it generalizes easily to non-likelihood models and is most easily robustified as detailed in Section 4.1. But it does have the limitation of lack of invariance to parameterization. For example, a test of $H_0 : \theta_1/\theta_2 = 1$ will lead in finite samples to Wald test statistic that differs from that for the equivalent hypothesis $H_0 : \theta_1 - \theta_2$. A bootstrap with asymptotic refinement, see section 4.3, should reduce this invariance.

The Lagrange multiplier or score test is less commonly-used in part because the usual method to compute these, by use of an auxiliary regression, has poor finite sample properties. Specifically, Monte Carlo studies find considerable over-rejection due to finite sample test size being considerably larger than the asymptotic size. A bootstrap with asymptotic refinement, however, can correct this problem. The LM test can be extended to non-likelihood settings, and can be robustified.

The likelihood ratio test generally does not extend to non-likelihood settings, though it does for optimal GMM estimation. Newey and West (1987b) generalize the three classical tests from the likelihood framework to the GMM framework.

The preceding hypothesis testing methods can also be used for model selection when models are nested. For model selection with nonnested models there is an extensive literature that we do not address here. A recent survey is provided by Pesaran and Weeks (2001).

In the remainder of this subsection we consider various model specification tests that do not rely on hypothesis tests of the form $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$.

The Hausman (1978) test contrasts two estimators that may be the same under a null hypothesis and differ under an alternative hypothesis. For example, one can compare OLS to the 2SLS estimator and conclude that there is endogeneity if the two estimators differ. Denote the two estimators by $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, in which case we test $H_0 : \text{plim}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) = \mathbf{0}$ using the statistic

$$H = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'(\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}])^{-1}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}],$$

which is chisquared distributed under H_0. Implementation requires estimating the variance matrix of the difference in the estimators. The original approach was to assume that one estimator, say $\widehat{\boldsymbol{\theta}}$ is efficient under the null, in which case $\mathrm{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}] = \mathrm{V}[\widetilde{\boldsymbol{\theta}}] - \mathrm{V}[\widehat{\boldsymbol{\theta}}]$. This is the standard method used today, even though it is generally incorrect since from Section 4.1 most applied studies use heteroskedastic-robust or cluster-robust standard errors that presume the estimator is in fact inefficient. One should instead use alternative methods to estimate $\mathrm{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}]$, such as the bootstrap.

Moment tests are tests of whether or not a population moment condition is supported by the data. So we test

$$\begin{aligned} H_0 &: \mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0} \\ H_a &: \mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] \neq \mathbf{0}. \end{aligned}$$

An obvious test is based on whether the corresponding sample moment $\widehat{\mathbf{m}} = N^{-1}\sum_i \mathbf{m}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$ is close to zero. The test statistic is

$$M = \widehat{\mathbf{m}}'(\widehat{\mathrm{V}}[\widehat{\mathbf{m}}])^{-1}\widehat{\mathbf{m}},$$

where M is chi-squared distributed under H_0 and the challenge is to estimate $\widehat{\mathrm{V}}[\widehat{\mathbf{m}}]$.

One leading example is an overidentifying restrictions (OIR) test. Then GMM estimation based on $\mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ cannot exactly impose $\widehat{\mathbf{m}} = \mathbf{0}$ if the model is overidentified. If GMM with optimal weighting matrix is used then Hansen (1982) showed that M is chisquared distributed under H_0 with degrees of freedom equal to the number of over-identifying restrictions.

A second class of examples are conditional moment tests where some model restrictions are used in estimation while other restrictions, not imposed in estimation, are used for specification testing. For example, in linear regression of y on $\mathbf{x}_1$ the hypothesis that $\mathbf{x}_2$ can be excluded as a regressor implies $\mathrm{E}[(y - \mathbf{x}_1'\boldsymbol{\beta}_1)|\mathbf{x}_2] = 0$ which can specified as a test of $H_0 : \mathrm{E}[(y - \mathbf{x}_1'\boldsymbol{\beta}_1)\mathbf{x}_2] = \mathbf{0}$. Here it can be difficult to obtain $\widehat{\mathrm{V}}[\widehat{\mathbf{m}}]$, though auxiliary regressions are available to compute an asymptotically equivalent

version of M in the special case that $\widehat{\boldsymbol{\theta}}$ is the MLE. Examples of conditional moment tests include the information matrix test of White (1982) and chisquared goodness-of-fit tests.

The Hausman test and OIR tests are routinely used in GMM applications. Conditional moment tests are less commonly used, even though they are easy to implement in likelihood settings and would seem especially useful then due to concerns of reliance on distributional assumptions. One reason is that the convenient auxiliary regressions used to compute then can have poor finite-sample size properties, but this can be rectified by a bootstrap with asymptotic refinement; see, for example, Horowitz (1996). A second reason is the more practical one that, especially with large samples, any model is quite likely to be rejected at conventional five percent significance levels.

4.3 Bootstrap

Inference in microeconometrics is based on asymptotic results that provide only an approximation given typical sample sizes. The bootstrap, introduced by Efron (1979), provides an alternative approximation by Monte Carlo simulation.

The motivation of the bootstrap is to view the data in hand or the fitted dgp as the population, draw B resamples from this population, and for each resample compute a relevant statistic. The empirical distribution of the resulting B statistics is used to approximate the distribution of the original statistic.

The most common use of the bootstrap is as a way to calculate standard errors. The data $\mathbf{w}_1, ..., \mathbf{w}_N$ are assumed to be iid distributed. The bootstrap standard error procedure is

1. Do the following B times:
   - Draw a bootstrap resample $\mathbf{w}_1^*, ..., \mathbf{w}_N^*$ by sampling with replacement from the original data (called a paired bootstrap).
   - Obtain estimate $\widehat{\theta}^*$ of θ, where for simplicity θ is scalar.
2. Use the B estimates $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$ to approximate the distribution of $\widehat{\theta}$. In particular, the bootstrap estimate of the standard error of $\widehat{\theta}$ is

$$s_{\widehat{\theta},\text{Boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \overline{\widehat{\theta}}^*)^2}, \tag{25}$$

where $\overline{\widehat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_b^*$. This is simply the standard deviation of $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$.

This method is convenient whenever standard errors are difficult to obtain by conventional methods. Leading examples are (1) two-step estimators when estimation at the first step complicates inference at the second step; (2) Hausman tests that require computation of the variance of the difference between two estimators when neither estimator is efficient under the null hypothesis; and (3) estimation with clustered errors when a package does not compute cluster-robust standard errors (in this case a cluster bootstrap that resamples over clusters is used). Given bootstrap standard errors, a standard Wald test of $H_0 : \theta = \theta_0$ uses $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta},\text{Boot}}$ and asymptotic normal critical values.

The preceding bootstrap is theoretically no better than usual first-order asymptotic theory. The attraction is the practical one of convenience.

Some bootstraps, however, provide a better asymptotic approximation, called an asymptotic refinement. The econometrics literature focuses on asymptotic refinement for test statistics. Consider a test of $H_0 : \theta = \theta_0$ with nominal significance level or nominal size α. An asymptotic approximation yields an actual rejection rate or true size $\alpha + O(N^{-j})$, where $O(N^{-j})$ means is of order N^{-j} and $j > 0$ with often $j = 1/2$ or $j = 1$. Then the true size goes to α as $N \to \infty$. Larger j is preferred, however, as then convergence to α is faster. A method with asymptotic refinement (or higher-order asymptotics) is one that yields j larger than that obtained using conventional asymptotics. The hope is that such asymptotic refinement will lead to tests with true size closer to α for moderate sample sizes, though this is not guaranteed. Asymptotic refinement may be possible if the bootstrap is applied to an asymptotically pivotal statistic, meaning one with asymptotic distribution that does not depend on unknown parameters.

The bootstrap standard error procedure does not lead to asymptotic refinement for the Wald test. Nor does the percentile method which rejects $H_0 : \theta = \theta_0$ if θ_0 falls outside the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$. The problem is that the bootstrap is of $\widehat{\theta}$ which is not asymptotically pivotal, since even under H_0 its asymptotic normal distribution depends on an unknown parameter (the variance).

Instead, the Wald statistic itself should be bootstrapped, as $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$ is asymptotically pivotal, since it is asymptotically $\mathcal{N}[0, 1]$ under H_0. The bootstrap-t or percentile-t procedure for a two-sided test of $H_0 : \theta = \theta_0$ at level α is

1. Do the following B times:

   - Draw a bootstrap resample $\mathbf{w}_1^*, ..., \mathbf{w}_N^*$ by sampling with replacement from the original data (called a paired bootstrap).
   - Obtain estimate $\widehat{\theta}^*$, standard error $s_{\widehat{\theta}^*}$ and t-statistic $t^* = (\widehat{\theta}^* - \widehat{\theta})/s_{\widehat{\theta}^*}$.

2. Use the B statistics $t_1^*, ..., t_B^*$ to approximate the distribution of $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$. For an equal-tailed (or nonsymmetrical) test reject H_0 if the original sample t-statistic falls outside the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates $t_1^*, ..., t_B^*$. For a symmetrical test reject H_0 if the original sample t-statistic falls outside the α quantile of $|t_1^*|, ..., |t_B^*|$.

Note that t^* in step 1 is centered on $\widehat{\theta}$ as the bootstrap views the original sample, with $\theta = \widehat{\theta}$, as the dgp. For equal-tailed two-sided tests (or for one-sided tests) this procedure leads to asymptotic refinement with true size $\alpha + O(N^{-1})$, rather than $\alpha + O(N^{-0.5})$ using bootstrap standard errors (or standard errors obtained using equation (21). For a two-sided symmetrical test (or a chisquared test) the corresponding rates are instead, respectively, $\alpha + O(N^{-1/2})$ and $\alpha + O(N^{-1})$.

There are many ways to bootstrap as there are different ways to obtain resamples, and there are many ways to use these resamples.

The resampling method used above is called a paired bootstrap as often $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ and here both y_i and $\mathbf{x}_i$ are being resampled. By contrast a residual bootstrap, for a model with additive error, holds $\mathbf{x}_i$ fixed and resamples over residuals $\widehat{u}_1, ..., \widehat{u}_N$ to yield resampled values $\mathbf{w}_i^* = (y_i^*, \mathbf{x}_i)$ where $y_i^* = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{u}_i^*$. A parametric bootstrap uses distributional knowledge, such as a specified distribution for $y_i|\mathbf{x}_i$ to resample. For hypothesis tests it is best, if possible, to impose H_0 in drawing the bootstrap sample.

Much development of the bootstrap has been done in the statistics literature. The econometrics literature is surveyed in Horowitz (2001), and MacKinnon (2002) provides much useful practical advice. Econometrics studies have focused on bootstraps for estimation methods used mainly by econometricians. For over-identified GMM models one should recenter so that the population moment condition is imposed in the sample. For non-smooth estimators and less than $\sqrt{N}$-consistent estimators standard bootstrap methods may provide inconsistent standard error estimates, leading to a currently active literature. Abrevaya and Huang (2005) consider the

maximum score estimator, Abadie and Imbens (2006b) consider matching treatment effects estimators, and Moreira, Porter, and Suarez (2004) consider IV with weak instruments. Subsampling, due to Politis and Romano (1994), works in a wider range of settings than the bootstrap.

In applied microeconometrics the main use of the bootstrap is to obtain standard errors. Bootstraps with asymptotic refinement are rarely done, as sample sizes are felt to be fairly large. But a bootstrap with asymptotic refinement can correct for many well-documented problems associated with standard tests, including the lack of invariance to parameterization for the Wald test and the poor finite-sample performance of auxiliary regressions used in computing Lagrange multiplier tests and conditional moment tests. And, in application with clustered observations, if there are few clusters there may be benefits to using a cluster bootstrap with asymptotic refinement.

5 Causation

The preceding sections present estimation and inference methods that can be used in a wide range of settings. Now we specialize to methods that can provide estimates of a causative effect, meaning measures of how an outcome changes in response to exogenous changes in a regressor. The "treatment effects" or "natural experiment" approach that extends randomized experiment methods to observational data, is presented in some detail Section 5.1. This major innovation in microeconometrics research uses a potential outcomes notation that differs from the simultaneous equations framework developed at the Cowles Commission. Developments in other more traditional methods to tease out causation, notably instrumental variables, panel data, and structural models, are presented in Sections 5.2 to 5.4. The final section presents partial identification which provides set identification (or bounds) under weaker assumptions than those necessary for point identification. The approach has much wider applicability than to causation, and is especially applicable to missing data problems.

5.1 Treatment Effects

We consider estimating the causal effect of a binary regressor. A stereotypical example is to consider the impact of a training program on earnings. The terminology of a medical trial is used. Then enrollment in a training program is viewed as treatment, having no training is viewed as control, and

we wish to estimate the causative effect of the treatment on the outcome variable earnings.

The ideal way to calculate this effect is to observe earnings for a person with the training, observe earnings for the same person without training, and subtract. But this is impossible. Instead we observe the outcome in only one state, while the other state is a hypothetical unobserved value, called a potential outcome or counterfactual.

The randomized experiment approach solves the inability to observe the counterfactual by comparing average outcomes, rather than individual outcomes, for two groups that are randomly assigned to either treatment or control. This approach is used at times in the social sciences, in social experiments. But most economics studies must instead rely on observational data.

The treatment effects literature seeks to extend the experimental approach to nonrandomized settings. Again averages across groups are compared, but now individuals select their treatment. Different assumptions about the nature of the self-selection of treatment and data availability lead to different methods to compute average effects of treatment.

The following framework is used. We consider a binary treatment, with variable d that takes value 1 if treatment is assigned and value 0 if untreated (a control). The observed outcome of interest y is a continuous variable that then takes value

$$y_i = \begin{cases} y_{1i} & \text{if treated } (d_i = 1) \\ y_{0i} & \text{if control } (d_i = 0). \end{cases} \tag{26}$$

The individual treatment effect is

$$\alpha_i = (y_{1i} - y_{0i}). \tag{27}$$

Note that (26) and (27) imply

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i} = y_{0i} + \alpha_i d_i. \tag{28}$$

Since only one of y_{1i} and y_{0i} are observed, α_i is not observable. Instead we try to estimate population averages of α_i, notably the average treatment effect (ATE)

$$\alpha_{\text{ATE}} = \text{E}[\alpha_i], \tag{29}$$

and the average treatment effect on the treated (ATET)

$$\alpha_{\text{ATET}} = \text{E}[\alpha_i | d_i = 1]. \tag{30}$$

These are conceptually quite different quantities. ATET gives the average gain in earnings for a person who actually receives training. ATE gives the

earnings gain averaged across those who did and those who did not receive the training.

The evaluation problem can be illustrated by decomposing ATET into two terms as

$$\alpha_{\text{ATET}} = \{\text{E}[y_{1i}|d_i = 1] - \text{E}[y_{0i}|d_i = 0]\} - \{\text{E}[y_{0i}|d_i = 1] - \text{E}[y_{0i}|d_i = 0]\}. \tag{31}$$

A naive estimate of α_{ATET} uses just the first term. But this ignores the second term, a selection term that arises if the treated and untreated are different in that on average they would have different untreated outcome. Methods differ according to whether this selection term can be solely controlled for by regressors, or whether it additionally depends on unobservables.

Given regressors $\mathbf{x}$, similar average effects can be defined, now varying with regressors. The average treatment effect is

$$\alpha_{\text{ATE}}(\mathbf{x}) = \text{E}[\alpha_i|\mathbf{X}_i = \mathbf{x}] \ . \tag{32}$$

and the average treatment effect on the treated

$$\alpha_{\text{ATET}}(\mathbf{x}) = \text{E}[\alpha_i|\mathbf{X}_i = \mathbf{x}, d_i = 1] \ . \tag{33}$$

Treatment effects are called heterogeneous if these quantities vary with the evaluation point $\mathbf{x}$, and are called homogeneous if $\alpha_{\text{ATE}}(\mathbf{x}) = \alpha_{\text{ATET}}(\mathbf{x}) = \alpha$. In practice results usually report estimates of the population measures $\alpha_{\text{ATE}} = \text{E}[\alpha_{\text{ATE}}(\mathbf{x})]$ and $\alpha_{\text{ATET}} = \text{E}[\alpha_{\text{ATET}}(\mathbf{x})]$ that average across individuals with different characteristics. For example, given individual-level estimates of $\widehat{\alpha}_{\text{ATE}}(\mathbf{x}_i)$ we can form $\widehat{\alpha}_{\text{ATE}} = N^{-1}\sum_{i=1}^{N}\widehat{\alpha}_{\text{ATE}}(\mathbf{x}_i)$.

We begin by assuming that selection is on observables only. Formally we make the conditional independence assumption that, conditional on regressors, outcomes are independent of treatment, so that

$$f(y_{ji}|\mathbf{x}_i, d_i = 1) = f(y_{ji}|\mathbf{x}_i, d_i = 0) = f(y_{ji}|\mathbf{x}_i),\ j = 0, 1. \tag{34}$$

This assumption of exogenous selection of treatment (given $\mathbf{x}$) is often written as $y_{0i}, y_{1i} \perp d_i \mid \mathbf{x}_i$ and has several other names including unconfoundedness and ignorability. For some purposes it can be weakened to apply to only y_{0i} or to apply only to conditional means (and not the entire distribution). The assumption implies that

$$\alpha_{\text{ATET}}(\mathbf{x}) = \text{E}[y_{1i}|\mathbf{X}_i = \mathbf{x}, d_i = 1] - \text{E}[y_{0i}|\mathbf{X}_i = \mathbf{x}, d_i = 0], \tag{35}$$

where the second term conditions on $d_i = 0$, rather than $d_i = 1$ as in the original definition (33).

The first method for estimating treatment effects is based on (35) and compares sample averages of y_1 and y_0 for individuals with the same level of $\mathbf{x}$. This **matching approach** permits treatment effects to be heterogeneous and provides nonparametric estimates of their average. In practice, however, such estimates become noisy or impossible as $\mathbf{x}$ will take many values if it is continuous or high-dimensional. One can instead use nonparametric methods such as kernel weighting that permit use of individuals with similar but not exactly the same level of $\mathbf{x}$. But more common is to match on the probability of treatment conditional on $\mathbf{x}$, or propensity score,

$$p(\mathbf{x}_i) = \Pr[d_i = 1|\mathbf{x}_i], \tag{36}$$

since Rosenbaum and Rubin (1983) showed that the conditional independence assumption carried over to conditioning on the propensity score (i.e., $y_{0i}, y_{1i} \perp d_i \mid p(\mathbf{x}_i)$). For example, nearest-neighbor propensity score matching uses

$$\widehat{\alpha}_{\text{ATET}} = N_1^{-1} \sum_{i:d_i=1} (y_{1i} - y_{0j}),$$

where $N_1 = \sum_{i=1}^{N} d_i$ and y_{0j} is the outcome for the nearest neighbor, the untreated observation with propensity score closest to that for y_{1i}. Other propensity score matching methods included kernel and stratification methods. The propensity score should be estimated using a flexible model such as a semiparametric binary model or a logit model with interactions. The propensity scores must have suitable common support over treatment and controls in order for matching to be feasible. For ATET we need $p(\mathbf{x}_i) < 1$, i.e. for any value of the regressors it is possible to not receive treatment, and for ATE we need $0 < p(\mathbf{x}_i) < 1$.

A second method is to specify and estimate a more restrictive regression model for the outcome. An obvious model is

$$y_i = \alpha d_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{37}$$

which imposes the constraint that the treatment effect α is homogeneous. OLS estimation of (37) yields a consistent estimate of the treatment effect α, assuming conditional independence and that (37) is correctly specified. This is called the **control function approach**, as the regressors $\mathbf{x}$ here include regressors that control for selection into treatment (i.e., explain d) as well as regressors that directly explain y in the absence of treatment. This more

parametric method has the advantage over matching of not requiring common support for the propensity score and permitting extrapolation beyond just the sample at hand.

The preceding two methods rely on the untestable assumption of conditional independence to be valid and presume that the data set is rich with many control variables, since observables alone are assumed sufficient to control for treatment selection. Should these conditions fail, which will be the case in many potential applications, the previous methods are invalid. For example, the OLS estimator in the simple homogeneous effects model (37) is inconsistent if the treatment indicator variable is correlated with the error term even after conditioning on regressors $\mathbf{x}$. Note also that even if treatment effects are heterogeneous and matching is valid, the estimates obtained are very problem specific and not necessarily generalizable to other settings.

A third method is to use **panel data fixed effects** estimators to control for unobserved heterogeneity. A panel data version of the homogeneous effects model (37) is

$$y_{it} = \alpha d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \phi_i + \delta_t + \varepsilon_{it}, \tag{38}$$

where here $\mathbf{x}_{it}$ does not include a constant and the intercept has both an individual-specific component ϕ_i and a time-specific component δ_t. We assume that treatment d_{it} is correlated with the unobservable ϕ_i, leading to inconsistency of OLS, but is uncorrelated with u_{it}. Then α can be consistently estimated by OLS estimation of the first-differences model

$$\Delta y_{it} = \alpha \Delta d_{it} + \Delta\delta_t + \Delta\mathbf{x}_{it}'\boldsymbol{\beta} + \Delta\varepsilon_{it}, \tag{39}$$

or by estimation of a mean-differences model (the fixed effect estimator), since ϕ_i has been eliminated. This standard method presumes panel data are available and is restricted to homogeneous treatment effects.

A fourth method, **differences in differences**, is applicable to repeated cross-sections, as well as panel data. We suppose there are just two periods, say $t = a$ (after) and $t = b$ (before), that all individuals are untreated in the first period and some are treated in the second period. Let $\bar{y}_{jt}$ denote the average outcome for treatment group $j = 0, 1$ in period $t = a, b$. The outcome changes over time by $(\bar{y}_{1a} - \bar{y}_{1b})$ in the treated group and by $(\bar{y}_{0a} - \bar{y}_{0b})$ in the untreated group. The differences in these differences provides an estimate of ATET, called the differences-in-differences estimator. This estimator is the OLS estimate of α in the model

$$y_{it} = \gamma + \alpha d_i + \beta e_t + u_{it},\ t = a, b,$$

where $d_i = 1$ is a binary treatment indicator and e_t is a binary time period indicator. Consistency of this estimator requires strong assumptions regarding the role of unobservables. In terms of (38) it is assumed that treatment selection does not depend on ε_{it} and that while it may depend on ϕ_i, on average $\text{plim}(\bar{\phi}_{ja} - \bar{\phi}_{jb}) = 0$. The method can be extended to estimate heterogeneous effects $\alpha_{\text{ATET}}(\mathbf{x})$ by grouping on $\mathbf{x}$ and then calculating within each group the four relevant averages of y.

A fifth method explicitly models the distribution of unobservables using **sample selection** models. These introduce a latent variable to explain treatment choice where the latent variable includes an unobserved component (or error) that is correlated with the error in the outcome equation. A linear model that permits heterogeneous effects and selection on unobservables is

$$\begin{aligned} y_{1i} &= \mathbf{x}_i'\boldsymbol{\beta}_1 + u_{1i} \\ y_{0i} &= \mathbf{x}_i'\boldsymbol{\beta}_0 + u_{0i} \\ d_i^* &= \mathbf{z}_i'\boldsymbol{\gamma} + v_i, \end{aligned} \tag{40}$$

where $d_i = 1$ if the latent variable $d_i^* > 0$, and $d_i = 0$ otherwise. A homogeneous effects version restricts $y_{1i} = y_{0i}$ aside from a difference of α in the intercept. Under the assumption that (u_{0i}, u_{1i}, v_i) are joint normal (with $\sigma_v^2 = 1$), some algebra yields

$$\begin{aligned} \mathrm{E}[y_{1i}|\mathbf{x}, d^* &> 0] = \mathbf{x}_i'\boldsymbol{\beta}_1 + \sigma_{1v}\lambda(\mathbf{z}_i'\boldsymbol{\gamma}), \\ \mathrm{E}[y_{0i}|\mathbf{x}, d^* &\leq 0] = \mathbf{x}_i'\boldsymbol{\beta}_0 - \sigma_{0v}\lambda(-\mathbf{z}_i'\boldsymbol{\gamma}), \end{aligned} \tag{41}$$

where $\lambda(\mathbf{z}'\boldsymbol{\gamma}) = \phi(\mathbf{z}'\boldsymbol{\gamma})/\Phi(\mathbf{z}'\boldsymbol{\gamma})$ is an inverse Mills ratio term and $\sigma_{jv} = \text{Cov}[u_{ji}, v_i]$. From (41) consistent estimates of $\boldsymbol{\beta}_1$ and σ_{1v} can be obtained by OLS estimation for the treated sample of y_1 on $\mathbf{x}$ and $\lambda(\mathbf{z}'\boldsymbol{\gamma})$, where $\widehat{\boldsymbol{\gamma}}$ is obtained by probit regression of d on $\mathbf{z}$. And OLS regression for the untreated sample of y_0 on $\mathbf{x}$ and $-\lambda(-\mathbf{z}'\boldsymbol{\gamma})$ gives consistent estimates of $\boldsymbol{\beta}_0$ and σ_{0v}. These estimates can then be used to estimate

$$\alpha_{\text{ATET}}(\mathbf{x}, \mathbf{z}) = \mathbf{x}_i'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\sigma_{0v} - \sigma_{1v})\lambda(\mathbf{z}_i'\boldsymbol{\gamma}).$$

The fundamental weakness of this sample selection approach is its reliance on distributional assumptions. These assumptions can be modified and relaxed, but even then the assumptions are still felt to be too strong.

A sixth method is **instrumental variables** estimation. Returning to the homogeneous effects model (37), the problem is that the regressor d is correlated with the error u. Assuming there is an instrument z that

does not belong in the model (so $\mathrm{E}[u|\mathbf{x}, z] = 0$) and is correlated with the treatment indicator d, the treatment effect α can be consistently estimated by IV regression of y on $\mathbf{x}$ and d with instruments $\mathbf{x}$ and z.

A seventh related method is estimation of the **local average treatment effects** (LATE). Begin with the homogeneous effects model with dependence on $\mathbf{x}$ dropped for simplicity, so

$$y_i = \beta + \alpha d_i + u_i. \tag{42}$$

Assume there is an instrument z with $\mathrm{E}[u|z] = 0$ and define $p(z) = \Pr[d = 1|Z = z] = \mathrm{E}[d|Z = z]$. Then

$$\mathrm{E}[y|z] = \beta + \alpha p(z).$$

Evaluating at two points z and z' and subtracting we obtain the local average treatment effect (LATE)

$$\alpha_{\mathrm{LATE}}(z) = \frac{\mathrm{E}[y|z] - \mathrm{E}[y|z']}{p(z) - p(z')}. \tag{43}$$

This can be estimated by comparing averages of the outcome y and treatment indicator d at two different values of the instrument z. If z is binary then this estimate is the same as the IV estimate. The estimate can be extended to heterogeneous effects, provided $p(z)$ is monotonic in z. Then it differs from IV and will vary with the points of evaluation z and z'. A more general treatment effect is the marginal treatment effect (MTE)

$$\alpha_{\mathrm{MTE}}(\mathbf{x}, z) = \left. \frac{\partial \mathrm{E}[y|\mathbf{x}, Z]}{\partial \Pr[d = 1|\mathbf{x}, Z]} \right|_{Z=z},$$

which gives the mean treatment effect for those at the margin of choosing treatment. ATE, ATET and LATE can be shown to be different weighted averages of MTE.

A final method is **regression discontinuity design**. We suppose treatment occurs when a variable s crosses a threshold $\bar{s}$, so that $d = 1(s > \bar{s})$, and the outcome y also depends on s. For example, a government program to improve school outcomes may be applied to schools in low income areas. A method is developed to calculate a score, and schools with score below a certain threshold receive the government program while those with higher score do not. A complication is that school outcome will directly depend on this score. The obvious approach is to compare y for those with s just less

than $\bar{s}$ to those with with s just greater than $\bar{s}$. But this will use only a small fraction of the data. Instead we use $\widehat{\alpha}$ from the least squares regression

$$y_i = \beta + \alpha d_i + \gamma h(s_i) + u_i, \tag{44}$$

where $h(\cdot)$ is a flexible function that is specified (e.g. polynomial) or is estimated by nonparametric methods. Given the discrete nature of the discontinuity at $\bar{s}$ it is clear that the method can also be used when effects are heterogeneous effects and will estimate ATE= $\mathrm{E}[\alpha_i|s_i]$ under mild additional assumptions. Another extension is to fuzzy designs where there is a discrete jump in treatment at $s = 1$, but this threshold is not sharp as some individuals with $s < \bar{s}$ are treated and some with $s > \bar{s}$ are untreated. Intuitively if a fraction f of the population in the immediate vicinity of $\bar{s}$ switch from untreated to treated then ATE is estimated by f times the estimated OLS coefficient of d in (44). This adaptation is qualitatively similar to that for LATE in (43).

The literature on treatment effects is vast. Econometricians have contributed to the literature on all the preceding methods, and the sample selection, IV and LATE methods originated in econometrics. Early econometrics papers, that generally did not explicitly use the current treatment effects framework, include Ashenfelter (1978), Heckman (1979), Heckman and Robb (1985), Lalonde (1986) and Björklund and Moffitt (1987). Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (1999) emphasize matching methods. Imbens and Angrist (1994) introduce LATE and Björklund and Moffitt (1987) and Heckman and Vytlacil (2000) introduce MTE. Van der Klaauw (2002) provides a detailed presentation of RD methods. More recent research provides distribution theory when a nonparametric component is used and seeks to extend methods to nonlinear models, for example Imbens and Athey (2006), and to multiple treatments. Brief surveys include Smith (2000), Blundell and Dias (2002) and Angrist (2006), while lengthier surveys include Heckman, Lalonde and Smith (1999) and Angrist and Krueger (1999).

5.2 Instrumental Variables Methods

We consider the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{45}$$

where $\mathrm{Cor}[\mathbf{x}_i, u_i] \neq \mathbf{0}$ so that OLS is inconsistent. Assume there exists an instrument $\mathbf{z}_i$ such that $\mathrm{Cor}[\mathbf{z}_i, u_i] = \mathbf{0}$. The IV estimator for a just-

identified model, considered for simplicity, is

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{46}$$

If $\text{Cor}[\mathbf{z}_i, u_i] = \mathbf{0}$ then $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and

$$\text{V}[\widehat{\boldsymbol{\beta}}_{\text{IV}}] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\Sigma\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}, \tag{47}$$

where $\Sigma = \text{E}[\mathbf{u}\mathbf{u}'|\mathbf{Z}]$. This estimator is easily extended to overidentified models, and to nonlinear models as a special case of GMM.

The applied literature has included many creative examples of instrument use. For example, in earnings-schooling regression a proposed instrument for schooling is distance to college, as this may be related to college attendance but may not directly effect earnings. Another possible instrument is birth month, which may be related to years of schooling as it determines age of school entry and hence years of schooling before a person reaches the minimum school leaving age.

This interest in use of IV methods has been somewhat diminished by recognition of problems that arise when instruments are weakly correlated with the regressor(s) being instrumented.

A weak instrument is one for which $\text{Cor}[\mathbf{z}_i, \mathbf{x}_i]$ is small. More precisely, suppose there is one endogenous regressor and several exogenous regressors. Then the instrument for the endogenous regressor is weak if the correlation between the endogenous regressor and the instrument is low after partialing out the effect of the other exogenous regressors. Then it is well-known that $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ will be imprecisely estimated. Two other complications can arise.

First, suppose that $\text{Cor}[\mathbf{z}_i, u_i]$ is close to zero rather than exactly zero. Then not only is the IV estimator inconsistent, but it can be more inconsistent than OLS estimator. For example, in the simple case of scalar regressor x and scalar instrument z, suppose the correlation between x and z is 0.1. Then IV becomes more inconsistent than OLS if the correlation between z and u exceeds a mere 0.1 times the correlation between x and u. This result, emphasized by Bound, Baker, and Jaeger (1995), has led to increased scrutiny of assumptions regarding the validity of an instrument in any particular application.

Second, even if $\text{Cor}[\mathbf{z}_i, u_i]$ equals zero, regular asymptotic theory performs poorly in finite samples if the instrument is weak. Theoreticians established key results early. Applied researchers to subsequently highlight the problem were Nelson and Startz (1990) and Bound, Jaeger and Baker (1995). Staiger and Stock (1997) provided influential theory.

There is a large theoretical literature on inference with weak instruments, including new testing procedures. Andrews and Stock (2005) provide a recent survey. A related literature considers inference when there are many instruments, meaning that as the sample size increases so too does the number of available instruments. A third area, IV estimation in the treatment effects setting, has already been mentioned.

5.3 Panel Data

Panel data are repeated observations on the same cross-section units, typically individuals or firms, for several time periods. The cross-section units are usually assumed to be independent, though this assumption may be less appropriate if the cross-section units are states or countries.

An obvious advantage of panel data is that they permit increased precision in estimation, due to an increased number of observations. It is important, however, that one control for likely correlation of observations over time for a given cross-section unit. The usual method is to use cluster-robust standard errors described in Section 4.1.

The microeconometrics literature has focused on a second advantage of panel data, that it provides a way to identify causation even if there is selection on unobservables, provided the unobservables are time-invariant.

The fixed effects linear panel model specifies

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},\ \ i = 1, ..., N, \tag{48}$$

where α_i and ε_{it} are unobserved. It is assumed that the idiosyncratic error ε_{it} is uncorrelated with $\mathbf{x}_{it}$, but the individual-specific error α_i is potentially correlated with $\mathbf{x}_{it}$. Note that while α_i is called a "fixed effect" in the literature, this term is misleading as it is being treated as random. We focus on short panels, with $N \to \infty$ but T permitted to be small (for a static linear model it is sufficient that $T \geq 2$).

To relate this to the treatment effects literature, $\mathbf{x}_{it}$ may include a binary treatment d_{it} that is correlated with the error term $\alpha_i + \varepsilon_{it}$ (selection on unobservables), but only with the component α_i of the error term that is time invariant. For example, an individual may self-select into a training program due to unobserved high ability, but this high ability is assumed to be time invariant.

Pooled OLS regression of y_{it} on $\mathbf{x}_{it}$ will lead to inconsistent estimation of $\boldsymbol{\beta}$, due to correlation of regressors with the error. The random effects estimator of $\boldsymbol{\beta}$, the feasible GLS estimator of (48) under the assumption that both α_i and ε_{it} are iid, is also inconsistent if in fact α_i is correlated

with $\mathbf{x}_{it}$. For this reason many microeconometrics studies shy away from random effects models that are widely used in other fields.

Estimation of transformed models that eliminate α_i can lead to consistent estimation. The fixed effects or within estimator is obtained by OLS estimation of the within model

$$(y_{it} - \bar{y}_i) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (u_{it} - \bar{u}_i). \tag{49}$$

The first-differences estimator is obtained by OLS estimation of the first-differences model

$$\Delta y_{it} = \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta u_{it}. \tag{50}$$

Note that in both cases only the coefficient of time-varying regressors can be identified.

Extension to nonlinear models is possible only for some specific models, as there is an incidental parameters problem. The asymptotics rely on $N \to \infty$, so the number of parameters (k regression coefficients plus N fixed effects α_i) is going to infinity with the sample size. Some models permit transformation that eliminate α_i, while others do not. For nonlinear models with additive error the within and first-differences transformations can be again used. For binary outcomes fixed effects estimation is possible for the logit model, see Chamberlain (1980), but not the probit model. For count data, Hausman, Hall and Griliches (1984) presented fixed effects estimation for the Poisson model and a particular parameterization of the negative binomial model. The Poisson fixed effects estimator does not require that the data be Poisson distributed, as it is consistent provided the conditional mean is correctly specified. An active area of research is developing methods for general nonlinear fixed effects panel models that while inconsistent due to the incidental parameters problem, are less inconsistent than existing methods. See, for example, Woutersen (2002).

Panel data also provide the opportunity to model individual-level dynamic behavior, since the individual is observed at more than one point in time. A simple dynamic linear fixed effects model includes a lagged dependent variable, so that

$$y_{it} = \rho y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},\ i = 1, ..., N. \tag{51}$$

An important result is that the fixed effects and first-differences estimators of this model are inconsistent. Instrumental variables estimation of the first-differences estimator is possible, using $y_{i,t-2}$ as an instrument for $(y_{i,t} - y_{i,t-1})$. Holtz-Eakin, Newey and Rosen (1988) and Arellano and Bond (1991)

proposed using additional lags as instruments and estimating by GMM using an unbalanced set of instruments.

For nonlinear dynamic models fixed effects estimation is possible for the logit model, see Chamberlain (1985) and Honore and Kyriazidou (2000), for the Poisson model, see Blundell, Griffiths and Windmeijer (2002), and for some duration models, see Chamberlain (1985) and Van den Berg (2001).

5.4 Structural Models

The classic linear simultaneous equations model (SEM) has deliberately not been discussed in this section on causation, as the SEM is rarely used in microeconometric studies. Many causal studies are interested in the marginal effect of a single regressor on a single dependent variable. In that case two-stage least squares regression of the single equation of interest is simply instrumental variables estimation, already discussed. And the IV estimator has deficiencies, leading to increased use of other methods given in this section. Finally the linear SEM does not extend readily to nonlinear models and in cases where it does, such as simultaneous equations Tobit models, the distributional assumptions are very strong.

Another type of structural modelling is microeconometrics models based on economic models of utility or profit maximization. Early references include Heckman (1974), MaCurdy (1981) and Dubin and McFadden (1984). The more recent labor literature most commonly uses structural economic models to explain employment dynamics, see for example Keane and Wolpin (1997). Reiss and Wolak (2005) provide a survey of structural modelling in industrial organization.

6 Heterogeneity

A loose definition of heterogeneity is that data differs across observations. In a regression context this heterogeneity may be due to regressors (observables) or due to unobservables.

We begin by considering heterogeneity due directly to observed regressors. For the linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $\mathrm{E}[u_i|\mathbf{x}_i] = 0$, $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ so that heterogeneity induces heterogeneity in the conditional means, though not in the marginal effects $\partial \mathrm{E}[y_i|\mathbf{x}_i]/\partial \mathbf{x}_i = \boldsymbol{\beta}$. Nonlinearity in the conditional mean, e.g. $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, will induce marginal effects that differ across individuals. Even simple parametric nonlinear models such as probit and Tobit have this feature. The standard method is to present a single summary statistic. Often the marginal effect is evaluated

at $\mathbf{x} = \bar{\mathbf{x}}$, but for most purposes a better single measure is the sample average of the individual marginal effects. Single-index models, i.e. $\mathrm{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i'\boldsymbol{\beta})$, have the advantage that the ratio of marginal effects for two different regressors equals the ratio of the corresponding parameters, and does not depend on the regressor values. So if one coefficient is twice another than the corresponding marginal effect is twice as large. Quite flexible modelling of heterogeneity in $\mathrm{E}[y_i|\mathbf{x}_i]$ and the associated marginal effects is possible using nonparametric regression of y on $\mathbf{x}$. This yields very noisy estimates for high dimension $\mathbf{x}$, leading to use instead of semiparametric methods such as those given in Section 3.6.

More challenging is controlling for unobserved heterogeneity that is due to factors other than the regressors. Then different individuals have different response even if the individuals have the same value of $\mathbf{x}$. Failure to control for this unobserved heterogeneity can lead to inconsistent parameter estimates and associated marginal effects. A simple example is omitted variables bias in the linear regression model, where the omitted variables form part of the unobserved heterogeneity. The source of the unobserved heterogeneity can also matter. In particular, in structural models of economic behavior a distinction is made as to whether or not the unobserved (to the econometrician) heterogeneity is known to the decision-maker.

Meaningful discussion of unobserved heterogeneity requires statement of an underlying structural relationship that we wish to estimate in the presence of unobserved heterogeneity. Wooldridge (2002, 2005) provides a fairly general framework. We suppose that interest lies in a conditional mean $m(\mathbf{x}, u) = \mathrm{E}[y|\mathbf{x}, u]$ or more formally

$$m(\mathbf{x}, u) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}, U = u],$$

where u is unobserved and for simplicity is a scalar. Ideally we would estimate $m(\mathbf{x}, u)$, but instead we are restricted to what Blundell and Powell (2004) call the average structural function (ASF)

$$m(\mathbf{x}) = \mathrm{E}_U[m(x, U)],$$

which integrates out the unobserved heterogeneity. Often interest lies in how ASF changes as the j^{th} regressor, say, changes. This is the average partial effect (APE)

$$\frac{\partial m(\mathbf{x})}{\partial x_j} = \mathrm{E}_U\left[\frac{\partial m(x, U)}{\partial x_j}\right].$$

Unobserved heterogeneity poses a problem because the ASF $m(\mathbf{x})$ in general differs from the conditional mean $\mathrm{E}[y|\mathbf{x}]$ $=\mathrm{E}_{U|\mathbf{x}}[m(x, U)]$, and hence APE

differs from $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j$, but it is only $\mathrm{E}[y|\mathbf{x}]$ that is identified from the observed data.

The simplest assumption, and one commonly made, is that u is independent of $\mathbf{x}$, as then $\mathrm{E}[y|\mathbf{x}] = m(\mathbf{x})$. In a model with additive heterogeneity analysis is particularly straightforward. If $m(\mathbf{x}, u) = g(\mathbf{x}, \boldsymbol{\beta}) + u$ then $\mathrm{E}[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta})$ given u independent of $\mathbf{x}$ with mean zero. Analysis is more complicated if unobserved heterogeneity enters in a nonlinear manner. For example, if $m(\mathbf{x}, u) = g(\mathbf{x}'\boldsymbol{\beta} + u)$ then $\mathrm{E}[y|\mathbf{x}] = \mathrm{E}_u[g(\mathbf{x}'\boldsymbol{\beta}+u)]$ will typically require specification of the distribution of u and integration over this. In some cases analytical expressions can be obtained. In other cases numerical methods are used. If u is low dimensional (in many applications it is a scalar) then Gaussian quadrature methods work well. Otherwise simulation methods given in Sections 4.3 and 4.4 can be used. Examples include negative binomial models for counts obtained by a Poisson-gamma mixture, Weibull-gamma mixtures for durations, random utility models for binary and multinomial data (where u is now a vector) and normal mixtures for linear and nonlinear panel data. While often the unobserved heterogeneity is interpreted as a random intercept, this can be generalized to random slopes (a random coefficients model). An alternative is finite mixtures models, used particularly in duration and count data analysis.

Panel data offer the opportunity to permit u to be dependent on $\mathbf{x}$. In that case u_{it} is decomposed into a time-varying component that is independent of $\mathbf{x}_{it}$ and a time-invariant component that may be correlated with $\mathbf{x}_{it}$. Fixed effects estimators for these models have been discussed in Section 5.3. It is important to note that in nonlinear models these methods identify $\boldsymbol{\beta}$ but not ASF, so that the APE's are only estimated up to scale.

Panel data also offer the possibility of distinguishing between persistence in behavior over time due to unobserved heterogeneity and persistent in behavior over time due to true state dependence. For example, rather than the static linear model $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$, where correlation of u_i with $\mathbf{x}_{it}$ causes problem, a more appropriate model may be a dynamic model $y_{it} = \rho y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$ where there is now no complication of unobserved heterogeneity. These models have quite different structural interpretation with quite different policy consequences. For example, high persistence of unemployment given regressors may be due to stigma attached to being unemployed (state dependence) or may be due to unobserved low ability (unobserved heterogeneity).

The treatment effects literature allows for unobserved heterogeneity. By assuming that selection is on observables it is possible to estimate ATET($\mathbf{x}$) which is the APE for the treatment variable.

Wooldridge (2002, 2005) proposes use of proxy variables to identify the ASF and APE. For simplicity, consider the linear model $y = \mathbf{x}'\boldsymbol{\beta} + u$. If $\mathrm{E}[u|\mathbf{x}] = 0$, then $m(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ so unobserved heterogeneity causes no problem. Now consider an omitted variables situation where $u = \mathbf{z}'\boldsymbol{\gamma}+\varepsilon$ with $\mathrm{E}[\varepsilon|\mathbf{x}] = 0$ but $\mathrm{E}[\mathbf{z}|\mathbf{x}] \neq \mathbf{0}$. The ASF is $m(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}+ \mathrm{E}[\mathbf{z}]'\boldsymbol{\gamma}$, whereas the conditional mean $\mathrm{E}[y|\mathbf{x}] = \boldsymbol{\beta}+ \mathrm{E}[\mathbf{z}|\mathbf{x}]'\boldsymbol{\gamma}$. These terms differ unless $\mathrm{E}[\mathbf{z}|\mathbf{x}] = \mathrm{E}[\mathbf{z}]$, the case if the unobserved heterogeneity is independent of $\mathbf{x}$. A weaker assumption than independence is to assume that there is a proxy variable $\mathbf{w}$ for $\mathbf{z}$ with the properties that (1) $\mathbf{x}$ and $\mathbf{z}$ are independent conditional on $\mathbf{w}$ so $\mathrm{E}[\mathbf{z}|\mathbf{x}, \mathbf{w}] = \mathrm{E}[\mathbf{z}|\mathbf{x}]$**,** and (2) $\mathrm{E}[y|\mathbf{x}, \mathbf{w}, \mathbf{z}, \varepsilon] = \mathrm{E}[y|\mathbf{x}, \mathbf{w}, \mathbf{z}, \varepsilon]$ so that $\mathbf{z}$ is redundant in the original model. Then $\mathrm{E}[y|\mathbf{x}, \mathbf{w}] = \mathbf{x}'\boldsymbol{\beta}+ \mathrm{E}[\mathbf{z}|\mathbf{w}]'\boldsymbol{\gamma}$ which can be identified by regression of y on $\mathbf{x}$ and $\mathbf{z}$. Taking the expected value with respect to $\mathbf{w}$ then gives the desired ASF. Wooldridge (2005) generalizes this approach to nonlinear models and argues that even though failure to control for unobserved heterogeneity may lead to inconsistent parameter estimates, it is still possible in some cases to consistently estimate ASF and APE.

There is also a growing literature on heterogeneity in nonparametric models. See for example, Matzkin (2006). A simple approach is to start with the conditional cdf $F(y|\mathbf{x})$ which can be nonparametrically estimated. If we define $u = F(y|\mathbf{x})$, then u is uniformly distributed on $(0, 1)$ and hence uncorrelated with $\mathbf{x}$. Inverting we obtain $y = F^{-1}(u|\mathbf{x}) = G(\mathbf{x}, u)$. This provides a decomposition into observables $\mathbf{x}$ and unobservables u that are independent of $\mathbf{x}$, a separable model. But this is not a structural model in the sense of the ASF given earlier.

Controlling for unobserved heterogeneity is an active area in microeconometrics, as much of the variation in the outcome is due to unobserved factors since typically $R^2 < 0.5$. It is particularly important when there is sample selection or self-selection. For example, in OLS regression we essentially require only that $\mathrm{E}[u|\mathbf{x}] = 0$, whereas if the sample is truncated or censored much stronger assumptions on u are needed even if semiparametric methods are used. Heckman (2000, 2005) and related papers explicitly considers heterogeneity and structural estimation. See also Blundell and Powell (2004) and Wooldridge (2005).

Still to come: IV and quantiles.

7 Data Issues

Microeconometrics data are usually survey data that come from sampling schemes more complicated than simple random sampling, and are often sub-

ject to measurement error or are even missing due to nonresponse. For completeness this section provides a survey of these topics, which receive relatively little attention in econometrics. Sampling schemes, measurement error and missing data are presented in, respectively, Sections 6.1 to 6.3.

7.1 Sampling Schemes

Survey data often use stratified and clustered sampling to lower interview costs and to provide more precise estimates for population subgroups, such as regions with relatively few people, than would otherwise be the case. The extensive sample survey literature, initially focused on estimation of population means but then extended to the regression case, has generally been ignored by the econometrics literature.

The first issue raised by survey sampling schemes is that the sample is no longer representative of the population.

If the sampling involves endogenous stratification, i.e., stratification is on a variable that is the dependent variable in a regression, then standard estimation methods lead to inconsistent parameter estimates. Examples include truncated regression (e.g., hours of work are modelled and only workers are surveyed), choice-based sampling (e.g., commute mode choice is modelled and bus-riders are deliberately oversampled as there are relative few bus riders), on-site sampling and case-control studies.

Let the conditional distribution of y given $\mathbf{x}$ be denoted $f(y|\mathbf{x},\boldsymbol{\theta})$. Usually the joint density of y and $\mathbf{x}$ is $g(y,\mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x},\boldsymbol{\theta}) \times g(\mathbf{x})$, where the parameters in $g(\mathbf{x})$ are suppressed. Since $g(\mathbf{x})$ does not involve $\boldsymbol{\theta}$, so that inference on $\boldsymbol{\theta}$ can be based on the conditional (on $\mathbf{x}$) log-likelihood. Under endogenous stratification, however, it can be shown that $g(y,\mathbf{x}|\boldsymbol{\theta})$ takes a more complicated form, and MLE needs to be based on the joint log-likelihood based on $g(y,\mathbf{x}|\boldsymbol{\theta})$, rather than only on $f(y|\mathbf{x},\boldsymbol{\theta})$.

Much of the econometrics literature has focused on choice-based sampling in discrete choice models, with references including Manski and Lerman (1977) who proposed a weighted MLE and Imbens (1992) who presented more efficient GMM methods. A more general presentation for endogenous stratification is given by Imbens and Lancaster (1996). Wooldridge (2002) considers inverse-probability weighted estimators.

If the sampling scheme involves stratification on exogenous regressors then the problems are less severe. Surveys provide sample weights that can be used to obtain population representative statistics. These sample weights need not be used in the typical situation where correct specification of a regression model is assumed. For example, if it is assumed that the

regression function is linear in $\mathbf{x}$, so that $y = \mathbf{x}'\boldsymbol{\beta} + u$ with $\mathrm{E}[u|\mathbf{x}] = 0$ so that $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then OLS is consistent even if the regressors $\mathbf{x}$ are not representative of the population in $\mathbf{x}$**.** If instead we wish to do OLS without the linearity assumption, then weighted OLS should be used as it provides an estimate of the so-called census coefficients; see DuMouchel and Duncan (1983). For example, a weighted OLS regression of earnings on years of schooling provides a consistent estimate of the population marginal effect on earnings of one more year of schooling, without assuming that the model is linear. Wooldridge (2001) gives a general treatment of weighted m-estimation.

By using knowledge of the exogenous stratification scheme it is possible to improve the efficiency of estimation, but this is rarely done in practice in part because the efficiency gains are felt to be relatively small.

The preceding discussion has presumed that observations are independent, but survey methods often induce dependence for subgroups of observations. For example, several households on the same block may be interviewed. Then data in that subgroup are likely to be positively correlated, and even after controlling for regressors, model errors are likely to be positively correlated.

Usually the assumption that errors are uncorrelated with regressors is maintained. The standard procedure is to use cluster-robust standard errors, presented in Section 4.1. Hierarchical linear models or multilevel models are often used in other social science disciplines. If clustering is felt to induce correlation of errors with regressors, then cluster-specific fixed effects, analogous to an individual-specific fixed effect in a panel data model, may also be used.

A related topic is that of spatial correlation where observations in nearby regions, such as adjoining states, are likely to be correlated. Such correlation is rarely controlled for, though methods have been developed. See, for example, Anselin (2001) and Lee (2004).

7.2 Measurement Error

Standard results for measurement error consider the linear regression model with classical measurement error. More recent work has considered nonlinear regression models and, in some cases, nonclassical measurement error.

Suppose $y = \beta x^* + u$, with error u uncorrelated with x^*, but we observe x rather than x^* and regress y on x. Then, from Angrist and Krueger (1999),

the OLS estimator $\widehat{\beta} = [\sum_i x_i^2]^{-1} \sum_i x_i y_i$ is in general inconsistent as

$$\begin{aligned} \operatorname{plim} \widehat{\beta} &= \left[\operatorname{plim} N^{-1} \sum\nolimits_i x_i^2\right]^{-1} \times \operatorname{plim} N^{-1} \sum\nolimits_i x_i(\beta x_i^* + u_i) \quad (52) \\ &= [\mathrm{V}[x]]^{-1}\mathrm{Cov}[x, x^*]\beta \\ &= \lambda\beta, \end{aligned}$$

where $\lambda = \mathrm{Cov}[x, x^*]/\mathrm{V}[x]$ is the reliability ratio of x as a measure of x^*, and we have assumed that $\operatorname{plim} N^{-1} \sum_i x_i u_i = 0$. This assumption that x is uncorrelated with u requires the additional assumption that u is uncorrelated with the measurement error $v = x - x^*$, in addition to the usual assumption that the model error u is uncorrelated with x^*.

The size of the inconsistency depends on the size of the reliability ratio, which has been measured in various survey validation studies. Angrist and Kruger (2001, p.1346) present a summary table with reliability ratios for log annual earnings, annual hours and years of schooling ranging from 0.71 to 0.94. Bound, Brown, and Mathiewetz (2001, pp. 3749-3830) summarize many validation studies for labor-related data that also indicate that measurement error is large enough to lead to appreciable bias in OLS coefficients.

Result (52) makes few assumptions beyond independence of measurement error and model error. Textbook treatments of measurement error emphasize the classical measurement error model, a more restrictive model, that assumes

$$\begin{aligned} y &= \beta x^* + u,\ u \sim iid\left[0, \sigma_u^2\right] \quad (53) \\ x &= x^* + v,\ v \sim iid\left[0, \sigma_v^2\right] \text{ and } x^* \sim iid\left[0, \sigma_{x^*}^2\right]. \end{aligned}$$

Then $\operatorname{plim} \widehat{\beta} = \lambda\beta$, where $\lambda = \sigma_{x^*}^2/(\sigma_{x^*}^2 + \sigma_v^2) = 1/(1+s)$ where $s = \sigma_v^2/\sigma_{x^*}^2$ is the noise-to-signal ratio. Since $s \geq 0$, $\widehat{\boldsymbol{\beta}}$ is downward biased asymptotically towards zero, a bias called attenuation bias. The attenuation bias is reduced if additional (correctly measured) regressors are included, and is increased if panel data are used with estimation by differencing methods such as the within estimator.

There are several ways to secure identification of $\boldsymbol{\beta}$. These include instrumental variables methods (assuming availability of an instrument z that is correlated with x^* but not with the model error u), use of replicated data or validation sample data to estimate key sample cross-moments, and use of additional distributional assumptions, such as symmetry of the error. Bounds on $\boldsymbol{\beta}$ can also be obtained using reverse regression. Wansbeek and Meijer (2000) review many identification methods.

The preceding methods do not generalize easily and in a systematic way to nonlinear models. Carroll, Ruppert and Stefanski (1995) summarize the statistics literature and Hausman (2001) considers the econometrics literature.

Y. Amemiya (1985), who focused on polynomial regression, showed that instrumental variables methods do not extend easily to nonlinear regression models with additive error. Most studies consider use of repeated measures; see, for example, Hausman, Newey and Powell (1995), Li (2002), and Schennach (2004). Schennach (2006) proposes an instrumental variables estimator.

For nonlinear models with nonadditive error, such as discrete outcome and count models, measurement error in either a regressor or the dependent variable cause problems. Hausman, Abrevaya and Scott-Morton (1998) consider mismeasurement in the dependent variable in binary outcome models. Guo and Li (2002) consider mismeasurement in a regressor in a Poisson model. These papers take a parametric approach with strong assumptions.

The classical measurement error maintains that the measurement error is iid. Some work relaxes this. An early example is that for binary regressor the measurement error is necessarily correlated with the true value, since the only way to mismeasure a value of 0 is as a 1, and vice-versa. Kim and Solon (2005) consider standard linear panel estimators when measurement error in a regressor is negatively correlated with the true value.

7.3 Missing Data

Missing data due to survey nonresponse is common. Simple corrections include dropping an observation if any variable is missing (listwise deletion or case deletion) and simple imputation methods such as using the sample average or predictions from a fitted regression model.

The modern approach is to use multiple imputation methods that regard missing data as random variables and replaces with draws from an assumed underlying distribution.

The starting point is terminology and assumptions made about the nature of the process leading to missing data on w_i, say, due to Rubin (1976). These have many similarities with the potential outcomes model where the unknown counterfactual can also be viewed as a missing data problem. If the probability of w_i being missing depends on neither its own value or on other data in the data set then w_i is missing completely at random (MCAR), and missing data on w_i causes no problems aside from efficiency loss. If the probability of w_i being missing depends on other data in the data set, but

not its own value, then w_i is missing completely at random (MAR), and missing data may lead to estimator inconsistency. If w_i is MAR, then it is adjust for missingness if the missing data mechanism is ignorable, meaning that the parameters of the missing data mechanism are unrelated to the parameters that we estimate, similar to weak exogeneity.

Let $\mathbf{W} = (\mathbf{W}_{obs}, \mathbf{W}_{miss})$ denote the data partitioned into observed and missing observations, and suppose $\mathbf{W}$ has density $f(\mathbf{W}|\boldsymbol{\theta})$. Suppose we obtain imputed value $\mathbf{W}_{miss}^{(I)}$ and then obtain the MLE based on $f(\mathbf{W}_{obs}, \mathbf{W}_{miss}^{(I)}|\boldsymbol{\theta})$. This will overstate estimator precision as it fails to account for the uncertainty created by imputation of $\mathbf{W}_{miss}^{(I)}$. Multiple imputation overcomes this by obtaining m different imputed values for $\mathbf{W}_{miss}$ and hence m estimates $\widehat{\boldsymbol{\theta}}_r$, $r = 1, ..., m$ with associated variance matrices $\widehat{\mathbf{V}}_r = \widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}_r]$. Then

The key is imputing $\mathbf{W}_{miss}$. The multiple imputation literature focuses on doing so assuming MAR with ignorable missingness. There are several ways to make imputations. A preferred, though computationally expensive method, is to use data augmentation and MCMC methods. Given an s^{th} round estimate of $\boldsymbol{\theta}^{(r)}$ we impute $\mathbf{W}_{miss}^{(r+1)}$ by making a draw from $f(\mathbf{W}_{miss}|\mathbf{W}_{obs}, \boldsymbol{\theta}^{(r)})$. Then a new estimate $\boldsymbol{\theta}^{(r+1)}$ is obtained by drawing from $f(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{miss}^{(r)})$. The chain is continued to convergence, giving an imputed value for $\mathbf{W}_{miss}$.

For further details see Little and Rubin (1987), Rubin (1987) and Schafer (1997).

Also bring in probability weighting and two-stage IV - Angrist and Krueger (1997) and Moffitt and Ridder (2006).

8 Conclusion

Microeconometricians are very ambitious in their desire to obtain marginal effects that can be given a causative interpretation, permit individual heterogeneity and are obtained under minimal assumptions. The associated statistical inference should also rely on minimal assumptions. This has led to a literature and toolkit that goes way beyond extending a linear structural equations model approach to a nonlinear setting.

This survey has of necessity been selective. The methods used in labor economics and public economics have been emphasized. General approaches have been presented, with specialization usually to the linear model. For econometrics methods for specific types of data - binary, multinomial, durations and counts - good starting points are the monographs by, respectively, Maddala (1983), Train (2002), Lancaster (1990), and Cameron and Trivedi

(1998), as well as the texts cited in the introduction.

9 References

Abadie, A., and G.W. Imbens (2006a), "Large Sample Properties of Matching Estimators for Average Treatment Effects," Econometrica, 235-267.

Abadie, A., and G.W. Imbens (2006b), "On the Failure of the Bootstrap for Matching Estimators," unpublished manuscript.

Abowd, J.M. and D. Card (1987), "On the Covariance of Earnings and Hours Changes," Econometrica, ??

Abrevaya, J., and J. Huang (2005), "On the Bootstrap of the Maximum Score Estimator," Econometrica, 1175-1204.

Altonji, J.G., and L.M. Segal (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14, 353-366.

Amemiya, T. (1985), Advanced Econometrics, Cambridge, MA, Harvard University Press.

Amemiya, Y. (1985), "Instrumental Variable Estimator for the Nonlinear Error in Variables Model," Journal of Econometrics, 28, 273-289.

Andrews, D.W.K., and M. Buchinsky (2000), "A Three-Step Method for choosing the Number of Bootstrap Replications," Econometrica, 68, 23-51.

Andrews, D.W.K., and J. Stock (2005), "Inference with Weak instruments," invited paper, 2005 World Congress of the Econometric Society.

Angrist, J., V. Chernozhukov, and I. Ferandez-Val (2006), "Quantile Regression Under Misspecification, with an Application to the U.S. Wage Structure," Econometrica, 539-563.

Angrist, J.D., and A.B. Krueger (1999), "Empirical Strategies in Labor Economics," in Handbook of Labor Economics, O.C. Ashenfelter and D.E. Card (Eds.), Volume 3A, 1277-1397, Amsterdam, North-Holland.

Angrist, J., and V. Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER Working Paper 9389.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Group Estimators," Oxford Bulletin of Economics and Statistics, 49, 431-434.

Arellano, M., and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," Review of Economic Studies, 58, 277-298.

Ashenfelter, 0. (1978), "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics, 60, 47-57.

Athey, S. and G.W. Imbens (2006), "Identification and Inference in Nonlinear Difference-in-Difference Models," Econometrica, 431-497.

Bellemare, C., B. Melenberg, and A. van Soest (2002), "Semiparametric Models for Satisfaction with Income," Portuguese Economic Journal, 1, 181-203.

Beran, R. (1982), "Estimating Sampling Distributions: The Bootstrap and Competitors," Annals of Statistics, 10. 212-225.

Bertrand, M., E. Duflo and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" Quarterly Journal of Economics, 119, 249-275.

Blundell, R., A. Gosling, H. Ichimura, C. Meghir (2007), "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition using Bounds," Econometrica, forthcoming.

Blundell, R., R. Griffith and F. Windmeijer (2002), "Individual Effects and Dynamics in Count Data Models," Journal of Econometrics, 102, 113-131.

Bound, J., C. Brown, and N. Mathiowetz (2001), "Measurement Error in Survey Data" in Handbook of Econometrics, J.J. Heckman and E.E. Leamer (Eds.), Volume 5, Amsterdam, North-Holland.

Bound, J., D.A. Jaeger, and R.M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," Journal of the American Statistical Association, 90, 443-450.

Buchinsky, M. (1994), "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," Econometrica, 62, 405-458.

Cameron, A.C., Gelbach, J., and D.L. Miller (2006a), "Bootstrap-Based Improvements for Inference with Clustered Errors," Working Paper No. 06-??, Department of Economics, University of California - Davis.

Cameron, A.C., Gelbach, J., and D.L. Miller (2006b), "Robust Inference with Multi-way Clustering," Working Paper No. 06-??, Department of Economics, University of California - Davis.

Cameron, A.C., and P.K. Trivedi (1998), Regression Analysis for Count Data, Econometric Society Monograph No. 30, Cambridge, UK, Cambridge University Press.

Cameron, A.C., and P.K. Trivedi (2005), Microeconometrics: Methods and

Applications, Cambridge, Cambridge University Press.

Carroll, R.J., D. Ruppert, and L.A. Stefanski (1995), Measurement Error in Nonlinear Models, London, Chapman and Hall.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," Review of Economic Studies, 47, 225-238.

Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5-46.

Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics*, Z. Griliches and M. Intrilligator (Eds.), Volume 2, 1247-1318, Amsterdam, North-Holland.

Chamberlain, G. (1985), "Heterogeneity, Omitted Variable Bias and Duration Dependence," in Longitudinal Analysis of Labor Market Data, J.J. Heckman and B. Singer (Eds.), 3-38, Cambridge, UK, Cambridge University Press.

Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," Econometrica, 245-262.

Chib, S. (1992), "Bayes Regression for the Tobit Censored Regression Model," Journal of Econometrics, 58, 79-99.

Chib, S. (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in J.J. Heckman and E.E. Leamer (Eds.), Handbook of Econometrics, Volume 5, 3570-3649, Amsterdam, North-Holland.

Davidson, R., and J.G. MacKinnon (1993), Estimation and Inference in Econometrics, Oxford, Oxford University Press.

Deb, P., and P.K. Trivedi (2002), "The Structure of Demand for Health Care: Latent Class versus Two-part Models," Journal of Health Economics, 21, 601-625.

Dehejia, R.H., and S. Wahba (1999), "Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association, 94, 1053-1062.

Donald, S. G., and Lang, K. (2004), "Inference with Differences in Differences and Other Panel Data", unpublished manuscript.

DuMouchel, W.K., and G.J. Duncan (1983), "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," Journal of the American Statistical Association, 78, 535-43.

Dubin, J.A., and D.L.McFadden (1984), "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," Econometrica, 52,

345-362.

Efron, B. (1979), "Bootstrapping Methods: Another Look at the Jackknife," Annals of Statistics, 7, 1-26.

Gelfland, A.E., and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," Journal of the American Statistical Association, 85, 398-409.

Geman, S., and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," Econometrica, 57, 1317-1339.

Geweke, J., G. Gowrisankaran, and R.J. Town (2003), "Bayesian Inference for Hospital Quality in a Selection Model," Econometrica, 71, 1215-1238.

Greene, W.H. (2003), Econometric Analysis, fifth edition, Upper Saddle River, NJ, Prentice-Hall.

Gouriéroux, C., and A. Monfort (1996), Simulation Based Econometrics Methods, New York, Oxford University Press.

Gouriéroux, C., A. Monfort, and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory," Econometrica, 52, 681-700.

Hahn, J., P. Todd and W. Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69, 201-209.

Haile and Tamer (2003), "Inference with an incomplete model of English auctions," Journal of Political Economy, 1-51.

Hall, P. (1992), The Bootstrap and Edgeworth Expansion, New York: Springer-Verlag.

Hall, P., and J.L. Horowitz (1996), "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," Econometrica, 64, 891-916.

Hansen, C. (2005), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," Journal of Econometrics, forthcoming .

Hansen, L.P. (1982), "Large Sample Properties of Generalized Methods of Moments Estimators," Econometrica, 1029-1054.

Hastings, W.K. (1970), "Monte Carlo Sampling Methods Using Markov Chain and Their Applications," Biometrika, 57, 97-109.

Hausman, J.A. (2001), "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," Journal of Economic Perspectives, 15, 57-68.

Hausman, J.A., J. Abrevaya, and F.M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete Response Setting," Journal of Econometrics, 87, 239-269.

Hausman, J.A., W.K. Newey, and J.L. Powell (1995), "Nonlinear Errors in Variables: Estimation of Some Engel Curves," Journal of Econometrics, 65, 205-233.

Heckman, J.J. (1974), "Shadow Prices, Market Wages, and Labor Supply," Econometrica, 42, 679-694.

Heckman, J.J. (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," Econometrica, 46, 931-960.

Heckman, J.J. (1979), "Sample Selection as a Specification Error," Econometrica, 47, 153-161.

Heckman, J.J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," Review of Economic Studies, 64, 605-654.

Holland, P.W. (1986), "Statistics of Causal Inference," Journal of the American Statistical Association, 81, 945-960.

Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988), "Estimating Vector Autoregressions with Panel Data," Econometrica, 56, 1371-1395.

Honore, B.E., and E. Kyriazidou (2000), "Panel Data Discrete Choice Models with Lagged Dependent Variables," Econometrica, 88, 839-874.

Horowitz, J. L. (1997), "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in Kreps and Wallis eds., Advances in Econometrics, Vol. 7.

Horowitz, J.L. (2001), "The Bootstrap," in Handbook of Econometrics, Volume 5, J.J. Heckman and E. Leamer (Eds.), 3159-3228, Amsterdam, North-Holland.

Huber, P.J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in Proceedings of the Fifth Berkeley Symposium, J. Neyman (Ed.), 1, 221-233, Berkeley, CA, University of California Press.

Imbens, G.W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," Econometrica, 60, 1187-1214.

Imbens, G.W. (2002), "Generalized Method of Moments and Empirical Likelihood," Journal of Business and Economic Statistics, 20, 493-506.

Imbens, G.W., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effect," Econometrica, 62, 467-475.

Imbens, G.W., and T. Lancaster (1996), "Efficient Estimation and Stratified Sampling," Journal of Econometrics, 74, 289-318.

Keane, M.P., and K.I. Wolpin (1997), "The Career Decisions of Young Men," Journal of Political Economy.

Kim, B, and G. Solon (2005), "Implications of Mean-reverting Measurement Error for Longitudinal Studies of Wages and Employment," Review of Economics and Statistics, 87, 193-196.

Kitamura, Y. (2006), "Empirical Likelihood Methods in Econometrics: Theory and Practice," Cowles Foundation Discussion Paper No. 1569.

Kloek, T., and H.K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," Econometrica, 46, 1-19.

Koenker, R. (2005), "Quantile Regression," Econometric Society Monograph, Cambridge, UK, Cambridge University Press.

Koenker, R., and G. Bassett (1978), "Regression Quantiles," Econometrica, 46, 33-50.

Koenker, R., and K.F. Hallock (2001), "Quantile Regression," Journal of Economic Perspectives, 15, 143-156.

Koop, G.M. (2003), Bayesian Econometrics, New York, Wiley.

Koop, G.M. , D.J. Poirier, and J.L. Tobias (2007), Bayesian Econometric Methods, Cambridge, Cambridge University Press.

Lancaster, T. (1990), The Econometric Analysis of Transitional Data, Cambridge, UK, Cambridge University Press.

Lancaster, T. (2004), An Introduction to Modern Bayesian Econometrics, Oxford, Blackwell.

Lerman, S.R., and C.F. Manski (1981), "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in Structural Analysis of Discrete Data with Econometric Applications, C.F. Manski, and D. McFadden (Eds.), 305-319, Cambridge, MA, MIT Press.

Li, T. (2002), "Robust and Consistent Estimation of Non-linear Errors-in-Variables Models," Journal of Econometrics, 110, 1-26.

Liang, K.-Y., and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," Biometrika, 73, 13-22.

Little, R.J.A., and D. Rubin (1987), Statistical Analysis with Missing Data, New York, John Wiley.

Ludwig, J., and D.L. Miller (2006),"Does Head Start Improve Children's Life Chances?: Evidence from a Regression Discontinuity Design," Quarterly Journal of Economics, forthcoming.

MacKinnon, J.G. (2002), "Bootstrap inference in Econometrics," Canadian Journal of Economics, 35, 615-645.

MaCurdy, T.E. (1981), "An Empirical Model of Labor Supply in a Life-Cycle Setting, Journal of Political Economy, 89, 1059-1085.

Maddala, G.S. (1983), Limited-Dependent and Qualitative Variables in Economics, Cambridge, UK, Cambridge University Press.

Mahajan, A. (2006), "Identification and Estimation of Regression Models with Misclassification," Econometrica, 631-665.

Manski, C.F. (1975), "The Maximum Score Estimator of the Stochastic Utility Model of Choice," Journal of Econometrics, 3, 205-228.

Manski, C.F., and S.R. Lerman (1977), "The Estimation of Choice Probabilities from Choice-Based Samples," Econometrica, 45, 1977-1988.

Manski, C.F. (1988a), Analog Estimation Methods in Econometrics, London, Chapman and Hall.

Manski, C.F. (1995), Identification Problems in the Social Sciences, Harvard University Press.

Manski, C.F. (2006), "Partial Identification in Econometrics," New Palgrave Dictionary of Economics, 2nd Edition,

Manski, C.F. and Pepper (2000), "Monotone Instrumental Variables with an Application to the Returns to Schooling," Econometrica, 997-1010.

Matzkin, R.L. (2005), "Identification in Nonparametric Simultaneous Equations," mimeo, Northwestern University.

Meyer, B.D. (1990), "Unemployment Insurance and Unemployment Spells," Econometrica, 58, 757-782.

McCullagh, P., and J.A. Nelder (1983, 1989), Generalized Linear Models, 1st and 2nd editions, London, Chapman and Hall.

McCulloch, R.E., N.G. Polson, and P.E. Rossi (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," Journal

of Econometrics, 2000

McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," Econometrica, 57, 995-1026.

Moulton, B.R. (1986), "Random Group Effects and the Precision of Regression Estimates," Journal of Econometrics, 32, 385-397.

Moulton, B.R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," Review of Economics and Statistics, 72, 334-38.

Moreira, M.J., J.R. Porter, and G.A. Suarez (2004), "Bootstrap and Higher-Order Expansion Validity When Instruments May Be Weak," NBER Technical Working Paper 302.

Nelson, C.R., and R. Startz (1990), "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One," Journal of Business, 63, S125-140.

Newey, W.K. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests," Econometrica, 53, 1047-1070.

Newey, W.K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

Newey, W.K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, R.F. Engle and D. McFadden (Eds.), Volume 4, 2111-2245, Amsterdam, North-Holland.

Newey, W.K., and R.J. Smith (2004), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," Econometrica, 219-255.

Newey, W.K., and K.D. West (1987), "Hypothesis Testing with Efficient Method of Moments Estimators," International Economic Review, 28, 777-787.

Owen, A.B. (1988), "Empirical Likelihood Ratios Confidence Intervals for a Single Functional," Biometrika, 75, 237-249.

Pagan, A.R., and A. Ullah (1999), Nonparametric Econometrics, Cambridge, UK, Cambridge University Press.

Pakes, A.S., and D. Pollard (1989), "Simulation and the Asymptotics of Optimization Estimators," Econometrica, 57, 1027-1057.

Politis, D.N., and J.P. Romano (1994), "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions," Annals of Statistics, 22, 2031-2050.

Powell, J.L. (1984), "Least Squares Absolute Deviations Estimation for the Censored Regression Model," Journal of Econometrics, 25, 303-325.

Powell, J.L. (1986), "Censored Regression Quantiles," Journal of Econometrics, 32, 143-155.

Qin, J., and J. Lawless (1994), "Empirical Likelihood and General Estimating Equations," Annals of Statistics, 22, 300-325.

Racine, J. S. and Q. Li (2004), "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," Journal of Econometrics, March, Volume 119, Issue 1, pp. 99-130.

Reiss, P.C., and F.A. Wolak (2005), Structural Econometric Modeling: Rationales and Examples from Industrial Organization, Handbook of Econometrics: volume 6, forthcoming.

Robinson, P.M. (1988), "Root-N-Consistent Semiparametric Regression," Econometrica, 56, 931-954.

Rosenbaum, P. and D.B. Rubin (1983), "The Central Role of Propensity Score in Observational Studies for Causal Effects," Biometrika, 70, 41-55.

Rubin, D.B. (1976), "Inference and Missing Data," Biometrika, 63, 581-592.

Rubin, D.B. (1978), "Bayesian Inference for Causal Effects," Annals of Statistics, 6, 34-58.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, New York, John Wiley.

Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data, London, Chapman and Hall.

Schennach, S. (2004), "Estimation of Nonlinear Models with Measurement Error," Econometrica, 72, 33-75.

Schennach, S. (2006), "Instrumental Variable Estimation in Nonlinear Errors-in-Variables Models," unpublished manuscript.

Severini, T.A., and G. Tripathi (2001), "A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models," *Journal of Econometrics*, 102, 23-66.

Staiger, D., and J. Stock (1997), "Instrumental Variables Regression with Weak Instruments," Econometrica, 65, 557-586.

Tauchen, G. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models," Journal of Econometrics, 30, 415-443.

Train, K.E. (2003), Discrete Choice Methods with Simulation, Cambridge,

UK, Cambridge University Press.

Van den Berg, G. (2001), "Duration Models: Specification, Identification, and Multiple Durations," in Handbook of Econometrics, Heckman, J.J., and E. Leamer (Eds.), Volume 5, 3381-3460, Amsterdam, North-Holland.

Wansbeek, T., and E. Meijer (2000), Measurement Error and Latent Variables in Econometrics, Amsterdam, North-Holland.

White, H. (1980a), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48, 817-838.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," Econometrica, 50, 1-25.

White, H. (1984, 200?), Asymptotic Theory for Econometricians, San Diego, Academic Press.

Windmeijer (2005), "A Finite Sample Correction for the Variance of Linear Two-Step GMM Estimators," Journal of Econometrics, 126, 25-51.

Wooldridge, J.M. (1991), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Variances," Journal of Econometrics, 47, 5-46.

Wooldridge, J.M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," Econometric Theory, 17, 451-470.

Wooldridge, J.M. (2003), "Cluster-Sample Methods in Applied Econometrics," American Economic Review, 93, 133-138.

Woutersen, T. (2002), "Robustness against Incidental Parameters," unpublished manuscript.