Preliminary. Comments invited

Fixed and Random Effects in Nonlinear Models

William Greene*

Department of Economics, Stern School of Business,
New York University,

January 29, 2001

Abstract

This paper surveys recently developed approaches to analyzing panel data with nonlinear models. We summarize a number of results on estimation of fixed and random effects models in nonlinear modeling frameworks such as discrete choice, count data, duration, censored data, sample selection, stochastic frontier and, generally, models that are nonlinear both in parameters and variables. We show that notwithstanding their methodological shortcomings, fixed effects are much more practical than heretofore reflected in the literature. For random effects models, we develop an extension of a random parameters model that has been used extensively, but only in the discrete choice literature. This model subsumes the random effects model, but is far more flexible and general, and overcomes some of the familiar shortcomings of the simple additive random effects model as usually formulated. Once again, the range of applications is extended beyond the familiar discrete choice setting. Finally, we draw together several strands of applications of a model that has taken a semiparametric approach to individual heterogeneity in panel data, the latent class model. A fairly straightforward extension is suggested that should make this more widely useable by practitioners. Many of the underlying results already appear in the literature, but, once again, the range of applications is smaller than it could be.

Keywords: Panel data, random effects, fixed effects, latent class, random parameters

JEL classification: C1, C4

* 44 West 4th St., New York, NY 10012, USA, Telephone: 001-212-998-0876; fax: 01-212-995-4218; e-mail: wgreene@stern.nyu.edu, URL www.stern.nyu.edu/~wgreene. This paper has benefited greatly from discussions with George Jakubson (more on this below) and Scott Thompson and from seminar groups at The University of Texas, University of Illinois, and New York University. Any remaining errors are my own.

1. Introduction

The linear regression model has been called the automobile of economics (econometrics). By extension, in the of analysis of panel data, the linear fixed and random effects models have surely provided most of the thinking on the subject. However, quite a bit of what is generally assumed about estimation of models for panel data is based on results in the linear model, such as the utility of group mean deviations and instrumental variables estimators, that do not carry over to nonlinear models such as discrete choice and censored data models. Numerous other authors have noted this, and have, in reaction, injected a subtle pessimism and reluctance into the discussion. [See, e.g., Hsiao (1986, 1996) and, especially, Nerlove (2000).] This paper will explore some of those differences and demonstrate that, although the observation is correct, quite natural, surprisingly straightforward extensions of the most useful forms of panel data results can be developed even for extremely complicated nonlinear models.

The contemporary literature on estimating panel data models that are outside the reach of the classical linear regression is vast and growing rapidly. Model formulation is a major issue and is the subject of book length symposia [e.g., much of Matyas and Sevestre (1996)]. Estimation techniques span the entire range of tools developed in econometrics. No single study could hope to collect all of them. The objective of this one is to survey a set of recently developed techniques that extend the body of tools used by the analyst in single equation, nonlinear models. The most familiar applications of these techniques are in qualitative and limited dependent variable models, but, as suggested below, the classes are considerably wider than that.

1.1. The Linear Regression Model with Individual Heterogeneity

The linear regression model with individual specific effects is

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + \varepsilon_{it},\ t = 1,...,T(i),\ i = i,...,N,$$

$$E[\varepsilon_{it}|\mathbf{x}_{i1},\mathbf{x}_{i2},...,\mathbf{x}_{iT(i)}] = 0,$$

$$\text{Var}[\varepsilon_{it}|\mathbf{x}_{i1},\mathbf{x}_{i2},...,\mathbf{x}_{iT(i)}] = \sigma^2.$$

Note that we have assumed the strictly exogenous regressors case in the conditional moments, [see Woolridge (1995)] and have not assumed equal sized groups in the panel. The vector $\boldsymbol{\beta}$ is a constant vector of parameters that is of primary interest, α_i embodies the group specific heterogeneity, which may be observable in principle (as reflected in the estimable coefficient on a group specific dummy variable in the fixed effects model) or unobservable (as in the group specific disturbance in the random effects model). Note, as well that we have not included time specific effects, of the form γ_t. These are, in fact, often used in this model, and our omission could be substantive. With respect to the fixed effects estimator discussed below, since the number of periods is usually fairly small, the omission is easily remedied just by adding a set of time specific dummy variables to the model. Our interest is in the more complicated case in which N is too large to do likewise for the group effects, for example in analyzing census based data sets in which N might number in the tens of thousands. For random effects models, we acknowledge that this omission might actually be relevant to a complete model specification. The analysis of two way models, both fixed and random effects, has been well worked out in the linear case. A full extension to the nonlinear models considered in this paper remains for further research. From this point forward, we focus on the common case of one way, group effect models.

1.2. Fixed Effects

The parameters of the linear model with fixed individual effects can be estimated by ordinary least squares. The practical obstacle of the large number of individual coefficients is overcome by employing the Frisch-Waugh (1933) theorem to estimate the parameter vector in parts. The "least squares dummy variable" (LSDV) or "within groups" estimator of $\boldsymbol{\beta}$ is computed by the least squares regression of $y_{it}^* = (y_{it} - \bar{y}_{i.})$ on the same transformation of $\mathbf{x}_{it}$ where the averages are group specific means. The individual specific dummy variable coefficients can be estimated using group specific averages of residuals, as seen in the discussion of this model in contemporary textbooks such as Greene (2000, Chapter 14). We note that the slope parameters can be estimated using simple first differences as well. However, using first differences induces autocorrelation into the resulting disturbance, so this produces a complication. [If $T(i)$ equals two, the approaches are the same.] Other estimators are appropriate under different specifications [see, e.g., Arellano and Bover (1995) and Hausman and Taylor (1981) who consider instrumental variables]. We will not consider these here, as the linear model is only the departure point, not the focus of this paper.

The fixed effects approach has a compelling virtue; it allows the effects to be correlated with the included variables. On the other hand, it mandates estimation of a large number of coefficients, which implies a loss of degrees of freedom. As regards estimation of $\boldsymbol{\beta}$, this shortcoming can be overstated. The typical panel of interest in this paper has many groups, so the contribution of a few degrees of freedom by each one adds to a large total. Estimation of α_i is another matter. The individual effects are estimated with the group specific data. That is, α_i is estimated with $T(i)$ observations. Since $T(i)$ might be small, and is, moreover, fixed, there is no argument for consistency of this estimator. Note, however, the estimator of α_i is inconsistent not because it estimates some other parameter, but because its variance does not go to zero in the sampling framework under consideration. This is an important point in what follows. In the linear model, the inconsistency of a_i, the estimator of α_i does not carry through into $\mathbf{b}$, the estimator of $\boldsymbol{\beta}$. The reason is that the group specific mean is a sufficient statistic; the incidental parameters problem is avoided. The LSDV estimator $\mathbf{b}_{LSDV}$ is not a function of the fixed effects estimators, $a_{i,LSDV}$.

1.3. Random Effects

The random effects model is a generalized linear model; if α_i is a group specific random disturbance with zero conditional mean and constant conditional variance, σ_α^2, then

$$\text{Cov}[\varepsilon_{it},\varepsilon_{is}|\, \mathbf{x}_{i1},\mathbf{x}_{i2},...,\mathbf{x}_{iT(i)}] = \sigma_\alpha^2 + \mathbf{1}(t = s)\sigma_\varepsilon^2 \;\; \forall\, t,s \mid i \text{ and } \forall\, i.$$

$$\text{Cov}[\varepsilon_{it},\varepsilon_{js}|\, \mathbf{x}_{i1},\mathbf{x}_{i2},...,\mathbf{x}_{iT(i)}] = 0 \;\; \forall\, t,s \mid i \neq j \text{ and } \forall\, i \text{ and } j.$$

The random effects linear model can be estimated by two step, feasible GLS. Different combinations of the residual variances from the linear model with no effects, the group means regression and the dummy variables produce a variety of consistent estimators of the variance components. [See Baltagi (1995).] Thereafter, feasible GLS is carried out by using the variance estimators to mimic the generalized linear regression of $(y_{it} - \theta_i \bar{y}_{i.})$ on the same transformation of $\mathbf{x}_{it}$ where $\theta_i = 1 - \{\sigma_\varepsilon^2/[T(i)\sigma_\alpha^2 + \sigma_\varepsilon^2]\}^{1/2}$. Once again, the literature contains vast discussion of alternative estimation approaches and extensions of this model, including dynamic models [see, e.g., Judson and Owen (1999)], instrumental variables [Arellano and Bover (1995)], and GMM estimation [Ahn and Schmidt (1995), among others in the same issue of the *Journal of*

Econometrics]. The primary virtue of the random effects model is its parsimony; it adds only a single parameter to the model. It's major shortcoming is its failure to allow for the likely correlation of the latent effects with the included variables - a fact which motivated the fixed effects approach in the first place.

1.4. Random Parameters

Swamy (1971) and Swamy and Arora (1972), and Swamy et. al. (1988a, b, 1989) suggest an extension of the random effects model to

$$y_{it} = \boldsymbol{\beta}_i'\mathbf{x}_{it} + \varepsilon_{it}, \ , \ t = 1,...,T(i), \ i = i,...,N$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i$$

where $E[\mathbf{v}] = \mathbf{0}$ and $Var[\mathbf{v}_i] = \boldsymbol{\Omega}$. By substituting the second equation into the first, it can be seen that this model is a generalized, groupwise heteroscedastic model. The proponents devised a generalized least squares estimator based on a matrix weighted mixture of group specific least squares estimators. This approach has guided much of the thinking about random parameters models, but it is much more restrictive than current technology provides. On the other hand, as a basis for model development, this formulation provides a fundamentally useful way to think about heterogeneity in panel data.

1.5. Modeling Frameworks

The linear model discussed above provides the benchmark for discussion of nonlinear frameworks. [See Matyas (1996) for a lengthy and diverse symposium.] Much of the writing on the subject documents the complications in extending these modeling frameworks to models such as the probit and logit models for binary choice or the biases that result when individual effects are ignored. Not all of this is so pessimistic, of course; for example, Verbeek (1990), Nijman and Verbeek (1992), Verbeek and Nijman (1992) and Zabel (1992) discuss specific approaches to estimating sample selection models with individual effects. Many of the developments discussed in this paper appear in some form in extensions of the aforementioned to binary choice and a few limited dependent variables. We will suggest numerous other applications below, and in Greene (2001). In what follows, several unified frameworks for nonlinear modeling with fixed and random effects and random parameters are developed in detail.

2. Nonlinear Models

We will confine attention at this point to nonlinear models defined by the density for an observed random variable, y_{it},

$$f(y_{it} \mid \mathbf{x}_{i1},\mathbf{x}_{i2},...,\mathbf{x}_{iT(i)}) = g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is a vector of ancillary parameters such as a scale parameter, or, for example in the Poisson model, an overdispersion parameter. As is standard in the literature, we have narrowed our focus to linear index function models, though the results below do not really mandate this; it is merely a convenience. The set of models to be considered is narrowed in other ways as well at this point. We will rule out dynamic effects; $y_{i,t-1}$ does not appear on the right hand side of the equation. (See, e.g., Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), Orme (1999), Heckman and MaCurdy(1980)]. Multiple equation models, such as VAR's are also left for later extensions. [See Holtz-Eakin (1988) and Holtz-Eakin, Newey and Rosen(1988, 1989).] Lastly, note that only the current data appear directly in the density for the current y_{it}. This is also a matter of convenience; the formulation of the model could be rearranged to relax this restriction with no additional complication. [See, again, Woolridge (1995).]

We will also be limiting attention to parametric approaches to modeling. The density is assumed to be fully defined. This makes maximum likelihood the estimator of choice.[1] Certainly non- and semiparametric formulations might be more general, but they do not solve the problems discussed at the outset, and they create new ones for interpretation in the bargain. (We return to this in the conclusions.) While IV and GMM estimation has been used to great advantage in recent applications,[2] our narrow assumptions have made them less attractive than direct maximization of the log likelihood. (We will revisit this issue below.)

The likelihood function for a sample of N observations is

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T(i)} g(y_{it}, \beta'\mathbf{x}_{it} + \alpha_i, \boldsymbol{\theta}),$$

How one proceeds at this point depends on the model, both for α_i (fixed, random, or something else) and for the random process, embodied in the density function. We will, as noted, be considering both fixed and random effects models, as well as an extension of the latter. Nonlinearity of the model is established by the likelihood equations,

$$\frac{\partial \log L}{\partial \beta} = \mathbf{0},$$

$$\frac{\partial \log L}{\partial \alpha_i} = 0, i = 1,..., N,$$

[1] There has been a considerable amount of research on GMM estimation of limited dependent and qualitative choice models. At least some of this, however, forces an unnecessarily difficult solution on an otherwise straightforward problem. Consider, for example, Lechner and Breitung (1996), who develop GMM estimators for the probit and tobit models with exogenous right hand side variables. In view of the results obtained here, in these two cases (and many others), GMM will represent an inferior estimator in the presence of an available, preferable alternative. (Certainly in more complicated settings, such as dynamic models, the advantage will turn the other way.)

[2] See, e.g., Ahn and Schmidt (1995) for analysis of a dynamic linear model and Montalvo (1997) for application to a general formulation of models for counts such as the Poisson regression model.

$$\frac{\partial \log L}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

which do not have explicit solutions for the parameters in terms of the data and must, therefore, be solved iteratively. In random effects cases, we estimate not α_i, but the parameters of a marginal density for α_i, $f(\alpha_i|\boldsymbol{\theta})$, where the already assumed ancillary parameter vector, $\boldsymbol{\theta}$, would include any additional parameters, such as the σ_α^2 in the random effects linear model.

We note before leaving this discussion of generalities that the received literature contains a very large amount of discussion of the issues considered in this paper, in various forms and settings. We will see many of them below. However, a search of this literature suggests that the large majority of the applications of techniques that resemble these is focused on two particular applications, the probit model for binary choice and various extensions of the Poisson regression model for counts. These two do provide natural settings for the applications for the techniques discussed here. However, our presentation will be in fully general terms. The range of models that already appear in the literature is quite broad. How broad is suggested by the list of already developed estimation procedures detailed in Appendix B.

3. Models with Fixed Effects

In this section, we will consider models which include the dummy variables for fixed effects. A number of methodological issues are considered first. Then, the practical results used for fitting models with fixed effects are laid out in full.

The log likelihood function for this model is

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i, \boldsymbol{\theta})$$

In principle, maximization can proceed simply by creating and including a complete set of dummy variables in the model. Surprisingly, this seems not to be common, in spite of the fact that although the theory is generally laid out in terms of a possibly infinite N, many applications involve quite a small, manageable number of groups. [Consider, for example, Schmidt, and Sickles' (1984) widely cited study of the stochastic frontier model, in which they fit a fixed effects linear model in a setting in which the stochastic frontier model would be wholly appropriate, using quite a small sample. See, as well, Cornwell, Schmidt, and Sickles (1990).] Nonetheless, at some point, this approach does become unusable with current technology. We are interested in a method that would accommodate a panel with, say, 50,000 groups, which would mandate estimating a total of 50,000 + K_β + K_θ parameters. That said, we will be suggesting just that. Looking ahead, what makes this impractical is a second derivatives matrix (or some approximation to it) with 50,000 rows and columns. But, that consideration is misleading, a proposition we will return to presently.

3.1. Methodological Issues in Fixed Effects Models

The practical issues notwithstanding, there are some theoretical problems with the fixed effects model. The first is the proliferation of parameters, just noted. The second is the 'incidental parameters problem.' Suppose that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ were known. Then, the solution for α_i would be based on only the $T(i)$ observations for group i. This implies that the asymptotic variance for a_i is $O[1/T(i)]$. Now, in fact, $\boldsymbol{\beta}$ is not known; it is estimated, and the estimator is a function of the estimator of α_i, $a_{i,ML}$. The asymptotic variance of $\mathbf{b}_{ML}$ must therefore be $O[1/T(i)]$ as well; the MLE of $\boldsymbol{\beta}$ is a function of a random variable which does not converge to a constant as $N \rightarrow \infty$. The problem is actually even worse than that; there is a small sample bias as well. The example is unrealistic, but in a binary choice model with a single regressor that is a dummy variable and a panel in which $T(i) = 2$ for all groups, Hsiao (1993, 1996) shows that the small sample bias is 100%. (Note, again, this is in the dimension of $T(i)$, so the bias persists even as the sample becomes large, in terms of N.) No general results exist for the small sample bias in more realistic settings. The conventional wisdom is based on Heckman's (1981) Monte Carlo study of a probit model in which the bias of the slope estimator in a fixed effects model was toward zero (in contrast to Hsiao) on the order of 10% when $T(i) = 8$ and $N = 100$. On this basis, it is often noted that in samples at least this large, the small sample bias is probably not too severe. Indeed, for many microeconometric applications, $T(i)$ is considerably larger than this, so for practical purposes, there is good cause for optimism. On the other hand, in samples with very small $T(i)$, the analyst is well advised of the finite sample properties of the MLE in this model.

In the linear model, using group mean deviations sweeps out the fixed effects. The statistical result at work is that the group mean is a sufficient statistic for estimating the fixed effect. The resulting slope estimator is not a function of the fixed effect, which implies that it (unlike the estimator of the fixed effect) is consistent. There are a number of like cases of

nonlinear models that have been identified in the literature. Among them are the binomial logit model,

$$g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) = \Lambda[(2y_{it}-1)(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)]$$

where $\Lambda(.)$ is the cdf for the logistic distribution. In this case, analyzed in detail by Chamberlain (1980), it is found that $\Sigma_t y_{it}$ is a sufficient statistic, and estimation in terms of the conditional density provides consistent estimator of $\boldsymbol{\beta}$. [See Greene (2000) for discussion.] Other models which have this property are the Poisson and negative binomial regressions [See Hausman, Hall, and Griliches (1984)] and the exponential regression model.

$$g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) = (1/\lambda_{it})\exp(-y_{it}/\lambda_{it}),\ \lambda_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i),\ y_{it} \geq 0.$$

[See Munkin and Trivedi (2000) and Greene (2001).] It is easy to manipulate the log likelihoods for these models to show that there is a solution to the likelihood equation for $\boldsymbol{\beta}$ that is not a function of α_i. Consider the Poisson regression model with fixed effects, for which

$$\log g(y_{it}, , \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) = -\lambda_{it} + y_{it}\log\lambda_{it} - \log y_{it}! \text{ where } \lambda_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i).$$

Write $\lambda_{it} = \exp(\alpha_i)\exp(\boldsymbol{\beta}'\mathbf{x}_{it})$. Then,

$$\log L = \sum_{i=1}^{N}\sum_{t=1}^{T(i)} -\exp(\alpha_i)\exp(\boldsymbol{\beta}'\mathbf{x}_{it}) + y_{it}(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) - \log y_{it}!$$

The likelihood function for α_i is

$$\frac{\partial \log L}{\alpha_i} = -\exp(\alpha_i)\sum_{t=1}^{T(i)}\exp(\boldsymbol{\beta}'\mathbf{x}_{it}) + \sum_{t=1}^{T(i)} y_{it} = 0.$$

The solution for α_i is given by

$$\exp(\alpha_i) = \frac{\sum_{t=1}^{T(i)} y_{it}}{\sum_{t=1}^{T(i)}\exp(\boldsymbol{\beta}'\mathbf{x}_{it})}.$$

This can be inserted back into the (now concentrated) log likelihood function where it can be seen that, in fact, the maximum likelihood estimator of $\boldsymbol{\beta}$ is not a function of α_i. The same exercise provides a similar solution for the exponential model.

There are other models, with linear exponential conditional mean functions, such as the gamma regression model. However, these are too few and specialized to serve as the benchmark case for a modeling framework. In the vast majority of the cases of interest to practitioners, including those based on transformations of normally distributed variables such as the probit and tobit models, this method of preceding will be unusable.

3.2. Computation of the Fixed Effects Estimator in Nonlinear Models

We consider, instead, brute force maximization of the log likelihood function, dummy variable coefficients and all. There is some history of this in the literature; for example, it is the approach taken by Heckman and MaCurdy (1980) and it is suggested quite recently by Sepanski (2000). It is useful to examine their method in some detail before proceeding. Consider the

probit model. For known set of fixed effect coefficients, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)'$, estimation of $\boldsymbol{\beta}$ is straightforward. The log likelihood conditioned on these values (denoted a_i), would be

$$\log L|\mathbf{a} = \sum_{i=1}^{N}\sum_{t=1}^{T(i)} \log g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + a_i)$$

This can be treated as a cross section estimation problem since with known $\boldsymbol{\alpha}$, there is no connection between observations even within a group. On the other hand, with given estimator of $\boldsymbol{\beta}$ (denoted $\mathbf{b}$) there is a conditional log likelihood function for each α_i, which can be maximized in isolation;

$$\log L_i|\mathbf{b} = \sum_{t=1}^{T(i)} \log \Phi[(2y_{it} - 1)(z_{it} + \alpha_i)]$$

where $z_{it} = \mathbf{b}'\mathbf{x}_{it}$ is now a known function. Maximizing this function (N times) is straightforward (if tedious, since it must be done for each i). Heckman and MaCurdy suggested iterating back and forth between these two estimators until convergence is achieved as a method of maximizing the full log likelihood function. We note three problems with this approach: First, there is no guarantee that this procedure will converge to the true maximum of the log likelihood function. The Oberhofer and Kmenta (1974) result that might suggest it would does not apply here because the Hessian is not block diagonal for this problem. Whether either estimator is even consistent in the dimension of N (that is, of $\boldsymbol{\beta}$) depends on the initial estimator being consistent, and there is no suggestion how one should obtain a consistent initial estimator. Second, in the process of constructing the estimator, the authors happened upon an intriguing problem. In any group in which the dependent variable is all 1s or all 0s, there is no maximum likelihood estimator for α_i - the likelihood equation for $\log L_i$ has no solution if there is no within group variation in y_{it}. This is an important feature of the model that carries over to the tobit model, as the authors noted. [See Maddala (1987) for further discussion.] A similar, though more benign effect appears in the loglinear models, Poisson and exponential and in the logit model. In these cases, any group which has $y_{it} = 0$ for all t contributes a 0 to the log likelihood function. As such, in these models as well, the group specific effect is not identified. Chamberlain (1980) notes this specifically; groups in which the dependent variable shows no variation cannot be used to estimate the group specific coefficient, and are omitted from the estimator. As noted, this is an important result for practitioners that will carry over to many other models. A third problem here is that even if the back and forth estimator does converge, even to the maximum, the estimated standard errors for the estimator of $\boldsymbol{\beta}$ will be incorrect. The Hessian is not block diagonal, so the estimator at the $\boldsymbol{\beta}$ step does not obtain the correct submatrix of the information matrix. It is easy to show, in fact, that the estimated matrix is too small. Unfortunately, correcting this takes us back to the impractical computations that this procedure sought to avoid in the first place.

Before proceeding to our 'brute force' approach, we note, once again, that data transformations such as first differences or group mean deviations are useless here.[3] The density is defined in terms of the raw data, not the transformation, and the transformation would mandate a transformed likelihood function that would still be a function of the nuisance parameters. 'Orthogonalizing' the data might produce a block diagonal data moment matrix, but it does not produce a block diagonal Hessian. We now consider direct maximization of the log likelihood function with all parameters. We do add one convenience. Many of the models we have studied

[3] This is true only in the parametric settings we consider. Precisely that approach is used to operationalize a version of the maximum score estimator in Manski (1987) and in the work of Honore (1992, 1996), Kyriazidou (1997) and Honore and Kyriazidou (2000) in the setting of censored data and sample selection models. As noted, we have limited our attention to fully parametric estimators.

involve an ancillary parameter vector, $\boldsymbol{\theta}$. However, no generality is gained by treating $\boldsymbol{\theta}$ separately from $\boldsymbol{\beta}$, so at this point, we will simply group them in the single parameter vector $\boldsymbol{\gamma} = [\boldsymbol{\beta}', \boldsymbol{\theta}']'$. It will be convenient to define some common notation: denote the gradient of the log likelihood by

$$\mathbf{g}_\gamma = \frac{\partial \log L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \frac{\partial \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \boldsymbol{\gamma}} \text{ (a } K_\gamma \times 1 \text{ vector)}$$

$$g_{\alpha i} = \frac{\log L}{\partial \alpha_i} = \sum_{t=1}^{T(i)} \frac{\partial \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \alpha_i} \text{ (a scalar)}$$

$$\mathbf{g}_\alpha = [g_{\alpha 1}, \ldots, g_{\alpha N}]' \text{ (an } N \times 1 \text{ vector)}$$

$$\mathbf{g} = [\mathbf{g}_\gamma', \mathbf{g}_\alpha']' \text{ (a } (K_\gamma + N) \times 1 \text{ vector).}$$

The full $(K_\gamma+N)\times(K_\gamma+N)$ Hessian is

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\gamma\gamma} & \mathbf{h}_{\gamma 1} & \mathbf{h}_{\gamma 2} & \cdots & \mathbf{h}_{\gamma N} \\ \mathbf{h}_{\gamma 1}' & h_{11} & 0 & \cdots & 0 \\ \mathbf{h}_{\gamma 2}' & 0 & h_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{h}_{\gamma N}' & 0 & 0 & 0 & h_{NN} \end{bmatrix}$$

where

$$\mathbf{H}_{\gamma\gamma} = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \text{ (a } K_\gamma \times K_\gamma \text{ matrix)}$$

$$\mathbf{h}_{\gamma i} = \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \boldsymbol{\gamma} \partial \alpha_i} \text{ (an } N \times 1 \text{ vector)}$$

$$h_{ii} = \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \alpha_i^{\,2}} \text{ (a scalar).}$$

Using Newton's method to maximize the log likelihood produces the iteration

$$\begin{pmatrix} \hat{\boldsymbol{\gamma}} \\ \hat{\alpha} \end{pmatrix}_k = \begin{pmatrix} \hat{\boldsymbol{\gamma}} \\ \hat{\alpha} \end{pmatrix}_{k-1} - \mathbf{H}_{k-1}^{-1} \mathbf{g}_{k-1} = \begin{pmatrix} \hat{\boldsymbol{\gamma}} \\ \hat{\alpha} \end{pmatrix}_{k-1} + \begin{pmatrix} \boldsymbol{\Delta}_\gamma \\ \boldsymbol{\Delta}_\alpha \end{pmatrix}$$

where subscript 'k' indicates the updated value and 'k-1' indicates a computation at the current value. We will now partition the inverse matrix. Let $\mathbf{H}^{\gamma\gamma}$ denote the upper left $K_\gamma \times K_\gamma$ submatrix of $\mathbf{H}^{-1}$ and define the $N \times N$ matrix $\mathbf{H}^{\alpha\alpha}$ and $K_\gamma \times N$ $\mathbf{H}^{\gamma\alpha}$ likewise. Isolating $\hat{\boldsymbol{\gamma}}$, then, we have the iteration

$$\hat{\gamma}_k = \hat{\gamma}_{k-1} - [\mathbf{H}^{\gamma\gamma}\mathbf{g}_\gamma + \mathbf{H}^{\gamma\alpha}\mathbf{g}_\alpha]_{k-1} = \hat{\gamma}_{k-1} + \Delta_\gamma$$

Now, we obtain the specific components in the brackets. Using the partitioned inverse formula [See, e.g., Greene (2000, equation 2-74)], we have

$$\mathbf{H}^{\gamma\gamma} = [\mathbf{H}_{\gamma\gamma} - \mathbf{H}_{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\gamma}]^{-1}.$$

The fact that $\mathbf{H}_{\alpha\alpha}$ is diagonal makes this computation simple. Collecting the terms,

$$\mathbf{H}^{\gamma\gamma} = \left[\mathbf{H}_{\gamma\gamma} - \sum_{i=1}^{N}\left(\frac{1}{h_{ii}}\right)\mathbf{h}_{\gamma i}\mathbf{h}_{\gamma i}'\right]^{-1}$$

Thus, the upper left part of the inverse of the Hessian can be computed by summation of vectors and matrices of order K_γ. We also require $\mathbf{H}^{\gamma\alpha}$. Once again using the partitioned inverse formula, this would be

$$\mathbf{H}^{\gamma\alpha} = \mathbf{H}^{\gamma\gamma}\mathbf{H}_{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1}$$

As before, the diagonality of $\mathbf{H}_{\alpha\alpha}$ makes this straightforward. Combining terms, we find that

$$\Delta_\gamma = -[\mathbf{H}^{\gamma\gamma}\mathbf{g}_\gamma + \mathbf{H}^{\gamma\alpha}\mathbf{g}_\alpha]_{k-1}$$

$$= -\mathbf{H}^{\gamma\gamma}(\mathbf{g}_\gamma - \mathbf{H}^{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{g}_\alpha)$$

$$= -\left[\mathbf{H}_{\gamma\gamma} - \sum_{i=1}^{N}\left(\frac{1}{h_{ii}}\right)\mathbf{h}_{\gamma i}\mathbf{h}_{\gamma i}'\right]^{-1}_{k-1}\left(\mathbf{g}_\gamma - \sum_{i=1}^{N}\frac{g_{\alpha i}}{h_{ii}}\mathbf{h}_{\gamma i}\right)_{k-1}$$

Turning now to the update for α, we use the like results for partitioned matrices. Thus,

$$\Delta_\alpha = -[\mathbf{H}^{\alpha\alpha}\mathbf{g}_\alpha + \mathbf{H}^{\alpha\gamma}\mathbf{g}_\gamma]_{k-1}.$$

Using Greene's (2-74) once again, we have

$$\mathbf{H}^{\alpha\alpha} = \mathbf{H}_{\alpha\alpha}^{-1}(\mathbf{I} + \mathbf{H}_{\alpha\gamma}\mathbf{H}^{\gamma\gamma}\mathbf{H}_{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1})$$

$$\mathbf{H}^{\alpha\gamma} = -\mathbf{H}^{\alpha\alpha}\mathbf{H}_{\alpha\gamma}\mathbf{H}_{\gamma\gamma}^{-1} = -\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\gamma}\mathbf{H}^{\gamma\gamma} \text{ (this is } -\mathbf{H}^{\gamma\alpha\prime})$$

Therefore,

$$\Delta_\alpha = -\mathbf{H}_{\alpha\alpha}^{-1}(\mathbf{I} + \mathbf{H}_{\alpha\gamma}\mathbf{H}^{\gamma\gamma}\mathbf{H}_{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1})\mathbf{g}_\alpha + \mathbf{H}_{\alpha\alpha}^{-1}(\mathbf{I} + \mathbf{H}_{\alpha\gamma}\mathbf{H}^{\gamma\gamma}\mathbf{H}_{\gamma\alpha}\mathbf{H}_{\alpha\alpha}^{-1})\mathbf{H}_{\alpha\gamma}\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{g}_\gamma.$$

After a bit of algebra, this reduces to

$$\Delta_\alpha = -\mathbf{H}_{\alpha\alpha}^{-1}(\mathbf{g}_\alpha + \mathbf{H}_{\alpha\gamma}\Delta_\gamma)$$

and, in particular, again owing to the diagonality of $\mathbf{H}_{\alpha\alpha}$

$$\boldsymbol{\Delta}_{\alpha i} = -\frac{1}{h_{ii}}\left(g_{\alpha i} + \mathbf{h}_{\gamma i}'\boldsymbol{\Delta}_{\boldsymbol{\gamma}}\right)$$

The important result here is that neither update vector requires storage or inversion of the $(K_\gamma+N)\times(K_\gamma+N)$ Hessian; each is computed as a function of sums of scalars and $K_\gamma\times 1$ vectors of first derivatives and mixed second derivatives.[4] The practical implication is that calculation of fixed effects models is a computation only of order K_γ and storage of N elements of $\boldsymbol{\alpha}$. Even for huge panels of hundreds of thousands of units, this is well within the capacity of even modest desktop computers of the current vintage. (We note in passing, the amount of computation is not particularly large either, though with the current vintage of 2+ GFLOP processors, computation time for econometric estimation problems is usually a nonissue.) One practical problem is that Newton's method is fairly crude, and in models with likelihood functions that are not globally concave, one might want to fine tune the algorithm suggested with a line search that prevents the parameter vector from straying off to proscribed regions in the early iterations.

This derivation concludes with the asymptotic variances and covariances of the estimators, which might be necessary later. For $\mathbf{c}$, the estimator of $\boldsymbol{\gamma}$, we already have the result we need. We have used Newton's method for the computations, so (at least in principle) the actual Hessian is available for estimation of the asymptotic covariance matrix of the estimators. The estimator of the asymptotic covariance matrix for the MLE of $\boldsymbol{\gamma}$ is $-\mathbf{H}^{\gamma\gamma}$, the upper left submatrix of $-\mathbf{H}^{-1}$. Note once again that this is a sum of $K_\gamma\times K_\gamma$ matrices which is of the form of a moment matrix and which is easily computed. Thus, the asymptotic covariance matrix for the estimated coefficient vector is easily obtained in spite of the size of the problem.

It is (presumably) not possible to store the asymptotic covariance matrix for the fixed effects estimators (unless there are relatively few of them). But, using the partitioned inverse formula once again, we can derive precisely the elements of Asy.Var[$\mathbf{a}$] that are contained in

$$\text{Asy.Var}[\mathbf{a}] = -[\mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\gamma}(\mathbf{H}_{\gamma\gamma})^{-1}\mathbf{H}_{\gamma\alpha}]^{-1}.$$

The ijth element of the matrix to be inverted is

$$(\mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\gamma}(\mathbf{H}_{\gamma\gamma})^{-1}\mathbf{H}_{\gamma\alpha})_{ij} = \mathbf{1}(i=j)h_{ii} - \mathbf{h}_{\gamma i}'(\mathbf{H}_{\gamma\gamma})^{-1}\mathbf{h}_{\gamma j}$$

This is a full $N\times N$ matrix, so the model size problem will apply - it is not feasible to manipulate this matrix as it stands. On the other hand, one could extract particular parts of it if that were necessary. For the interested practitioner, the Hessian to be inverted for the asymptotic covariance matrix of $\mathbf{a}$ is

$$\mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\gamma}(\mathbf{H}_{\gamma\gamma})^{-1}\mathbf{H}_{\gamma\alpha}$$

We keep in mind that $\mathbf{H}_{\alpha\alpha}$ is an $N\times N$ diagonal matrix. Using result 2-66b in Greene (2000), we have that the inverse of this matrix is

[4] This result appears in terse form in the context of a binary choice model in Chamberlain (1980, page 227). A formal derivation of the result was given to the author by George Jakubson of Cornell University in an undated memo, "Fixed Effects (Maximum Likelihood) in Nonlinear Models" with a suggestion that the result should prove useful for current software developers. (We agree.) Concurrent discussion with Scott Thompson at the Department of Justice contributed to this development.

$$[\mathbf{H}_{\alpha\alpha}]^{-1} + [\mathbf{H}_{\alpha\alpha}]^{-1}\mathbf{H}_{\alpha\gamma}\{(\mathbf{H}_{\gamma\gamma})^{-1} - \mathbf{H}_{\gamma\alpha}[\mathbf{H}_{\alpha\alpha}]^{-1}\mathbf{H}_{\alpha\gamma}\}^{-1}\mathbf{H}_{\gamma\alpha}[\mathbf{H}_{\alpha\alpha}]^{-1}.$$

By expanding the summations where needed, we find

$$Asy.Cov[a_i, a_j] = \mathbf{1}(i=j)\frac{1}{h_{ii}} + \frac{1}{h_{ii}}\frac{1}{h_{jj}}\mathbf{h}_{\gamma i}'\left[\mathbf{H}_{\gamma\gamma'}^{-1} - \sum_{i=1}^{N}\frac{1}{h_{ii}}\mathbf{h}_{\gamma i}\mathbf{h}_{\gamma i}'\right]^{-1}\mathbf{h}_{\gamma j}$$

Once again, the only matrix to be inverted is $K_\gamma \times K_\gamma$, not $N \times N$ (and, it is already in hand) so this can be computed by summation. It involves only $K_\gamma \times 1$ vectors and repeated use of the same $K \times K$ inverse matrix. Likewise, the asymptotic covariance matrix of the slopes and the constant terms can be arranged in a computationally feasible format. Using what we already have and result (2-74) in Greene (2000), we find that

$$Asy.Cov[\mathbf{c},\mathbf{a}'] = -\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\alpha} \times \text{Asy.Var}[\mathbf{a}].$$

Once again, this involves $N \times N$ matrices, but it simplifies. Using our previous results, we can reduce this to

$$Asy.Cov[\mathbf{c},a_i] = -\mathbf{H}_{\gamma\gamma}^{-1}\sum_{m=1}^{N}\mathbf{h}_{\gamma i}Asy.Cov[a_i, a_m].$$

This asymptotic covariance matrix involves a large amount of computation, but essentially no computer memory - only the $K_\gamma \times K_\gamma$ matrix. The $K_\gamma \times 1$ vectors would have to be computed 'in process,' which is why this involves a large amount of computation. At no point is it necessary to maintain an $N \times N$ matrix, which has always been viewed as the obstacle. Finally, we note the motivation for the last two results. One might be interested in the computation of an asymptotic variance for a function $g(\mathbf{b}, a_i)$ such as a prediction or a marginal effect for a probit model which has conditional mean function $\Phi(\mathbf{b}'\mathbf{x}_{it} + a_i)$. The delta method would require a very large amount of computation, but it is feasible with the preceding results.

A significant omission from the preceding is the nonlinear regression model. But, extension of these results to nonlinear least squares estimation of the model

$$y_{it} = f(\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i) + \varepsilon_{it}$$

is trivial. By defining the criterion function to be

$$\log L = -\sum_{i=1}^{N}\sum_{t=1}^{T(i)}\frac{1}{2}\varepsilon_{it}^2$$

all of the preceding results apply essentially without modification. Nonlinear least squares is often differentiated from other optimization problems. Jakubson suggests an alternative interpretation based on the Gauss-Newton regression on first derivatives only. In this iteration, the update vector is

$$\begin{pmatrix}\Delta_\beta \\ \Delta_\alpha\end{pmatrix} = (\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{e}$$

where **G** is the matrix with it'th row equal to the derivative of the conditional mean with respect to the parameters (i.e., the 'pseudo-regressors') and **e** is the current vector of residuals. As Jakubson

notes, with a minor change in notation, this computation is identical to the optimization procedure described earlier. (E.g., the counterpart to $\mathbf{H}_{\gamma\gamma}$ in this context will be $\mathbf{G}_{\gamma}'\mathbf{G}_{\gamma}$.)

With the exceptions noted earlier (binomial logit, Poisson and negative binomial - the exponential appears not to have been included in this group, perhaps because applications in econometrics have been lacking) the fixed effects estimator has seen relatively little use in nonlinear models. The methodological issues noted above have been the major obstacle, but the practical difficulty seems as well to have been a major deterrent. For example, after a lengthy discussion of a fixed effects logit model, Baltagi (1995) notes that "... the probit model does not lend itself to a fixed effects treatment." In fact, the fixed effects probit model is one of the simplest applications listed in the Appendix. (We note, citing Greene (1993), Baltagi (1995) also remarks that the fixed effects logit model as proposed by Chamberlain (1980) is computationally impractical with $T > 10$. This (Greene) is also incorrect. Using an extremely handy result from Krailo and Pike (1984), it turns out find that Chamberlain's binomial logit model is quite practical with $T(i)$ up to as high as 100. Consider, as well, Maddala (1987) who states

> "By contrast, the fixed effects probit model is difficult to implement computationally. The conditional ML method does not produce computational simplifications as in the logit model because the fixed effects do not cancel out. This implies that all *N* fixed effects must be estimated as part of the estimation procedure. Further, this also implies that, since the estimates of the fixed effects are inconsistent for small *T*, the fixed effects probit model gives inconsistent estimates for β as well. Thus, in applying the fixed effects models to qualitative dependent variables based on panel data, the logit model and the log-linear models seem to be the only choices. However, in the case of random effects models, it is the probit model that is computationally tractable rather than the logit model." (Page 285)

While the observation about the inconsistency of the probit fixed effects estimator remains correct, as discussed earlier, none of the other assertions in this widely referenced source are correct. The probit estimator is actually extremely easy to compute. Moreover, the random effects logit model is no more complicated than the random effects probit model. (One might surmise that Maddala had in mind the lack of a natural mixing distribution for the heterogeneity in the logit case, as the normal distribution is in the probit case. The mixture of a normally distributed heterogeneity in a logit model might seem unnatural at first blush. However, given the nature of 'heterogeneity' in the first place, the normal distribution as the product of the aggregation of numerous small effects seems less ad hoc.) In fact, the computational aspects of fixed effects models for many models are not complicated at all. We have implemented this model in over twenty different modeling frameworks including discrete choice models, sample selection models, stochastic frontier models, and a variety of others. A partial list appears in Appendix B to this paper.

4. Random Effects and Random Parameters Models

The general form of a nonlinear random effects model would be

$$f(y_{it}|\mathbf{x}_{it},u_i) = g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it}, u_i, \boldsymbol{\theta})$$

where

$$f(u_i) = h(u_i|\boldsymbol{\theta}).$$

Once again, we have focused on index function models, and subsumed the parameters of the heterogeneity distribution in $\boldsymbol{\theta}$. We do not assume immediately that the random effect is additive or that it has zero mean. As stated, the model has a single common random effect, shared by all observations in group i. By construction, the $T(i)$ observations in group i are correlated and jointly distributed with a distribution that does not factor into the product of the marginals. An important step in the derivation is the assumption at this point that conditioned on u_i, the $T(i)$ observations are independent. (Once again, we have assumed away any dynamic effects.) Thus, the joint distribution of the $T(i)$+1 random variables in the model is $f(y_{i1},y_{i2},...,y_{iT(i)},u_i \mid \mathbf{x}_{i1},...,\boldsymbol{\beta},\boldsymbol{\theta})$ which can be written as the product of the density conditional on u_i times $f(u_i)$;

$$f(y_{i1},y_{i2},...,y_{iT(i)},u_i \mid \mathbf{x}_{i1},...,\boldsymbol{\beta},\boldsymbol{\theta}) = f(y_{i1},y_{i2},...,y_{iT(i)}, \mid \mathbf{x}_{i1}, ..., u_i, \boldsymbol{\beta},\boldsymbol{\theta})\, f(u_i)$$

$$= \prod_{t=1}^{T(i)} g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it}, u_i, \boldsymbol{\theta})\, h(u_i|\boldsymbol{\theta})$$

In order to form the likelihood function for the observed data, u_i must be integrated out of this. With this assumption, skipping a step in the algebra, we obtain the log likelihood function for the observed data,

$$\log L = \sum_{i=1}^{N} \log\left[\int_{u_i} \left(\prod_{t=1}^{T(i)} g(y_{it}, \beta'\mathbf{x}_{it}, u_i, \theta) \right) h(u_i \mid \theta) du_i \right]$$

Three broadly defined approaches have been used to maximize this kind of likelihood.

4.1. Exact Integration and Closed Forms

In a (very) few cases, the integral contained in square brackets has a closed form which leaves a function of the data, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, which is then maximized using conventional, familiar techniques. Hausman, Hall and Griliches' (1984) analysis of the Poisson regression model is a widely cited example. If

$$f(y_{it}|\mathbf{x}_{it}, u_i) = \exp(-\lambda_{it}|u_i)(\lambda_{it}|u_i)^{y_{it}} / y_{it}!,\ \lambda_{it}|u_i = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + u_i)$$

where $v_i = \exp(u_i)$ has a gamma density with mean 1,

$$h(v_i|\theta) = \frac{\theta^{\theta}}{\Gamma(\theta)} e^{-\theta v_i} v_i^{\theta-1}, v \geq 0, \theta > 0$$

then the unconditional joint density for $(y_{i1},...,y_{iT(i)}|\mathbf{x}_{i1},...,\mathbf{x}_{iT(i)})$ has a negative binomial form

$$f(y_{i1},\ldots,y_{i,T(i)}) = \frac{\left(\prod_{t=1}^{T(i)} \lambda_{it}^{\ y_{it}}\right) \Gamma\left(\theta + \sum_{t=1}^{T(i)} y_{it}\right)}{\Gamma(\theta)\left(\prod_{t=1}^{T(i)} y_{it}!\right)\left[\left(\sum_{t=1}^{T(i)} y_{it}\right)!\right]\left(\sum_{t=1}^{T(i)} \lambda_{it}\right)^{\sum_{t=1}^{T(i)} y_{it}}} w_i^{\theta}(1-w_i)^{\sum_{t=1}^{T(i)} y_{it}}$$

The authors also obtained a closed form for the negative binomial model with log-gamma heterogeneity.

Finally, the stochastic frontier model is widely used framework in which the random effects model has a closed form. The structure of the most widely employed variant of the model is

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it} - |U_i|$$

where

$$v_{it} \sim N[0,\sigma_v^2]$$

$$U_i \sim N[0,\sigma_u^2].$$

(Note that the absolute value of the random effect appears in the index function model.) In this model, the mixture of normal distributions produces a fairly straightforward functional form for the log likelihood. The log likelihood function for this model was derived by Pitt and Lee (1981); the contribution of the ith group (observation) is

$$\text{Log } L_i = \frac{-\log 2\pi}{2} - \frac{(T(i)-1)\log\sigma_v^2}{2} - \frac{\log\left(\sigma_v^2 + T(i)\sigma_u^2\right)}{2} + \log\Phi\left(\frac{\mu_i^*}{\sigma_i^*}\right) - \frac{1}{2\sigma_v^2}\sum_{t=1}^{T(i)}\varepsilon_{it}^2 + \frac{1}{2}\left(\frac{\mu_i^*}{\sigma_i^*}\right)$$

where

$$\varepsilon_{it} = y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it}$$

$$\mu_i^* = -\frac{\sigma_u^2 \sum_{t=1}^{T(i)} \varepsilon_{it}}{\sigma_v^2 + T(i)\sigma_u^2}$$

$$\sigma_i^* = \frac{\sigma_u\sigma_v}{\sqrt{\sigma_v^2 + T(i)\sigma_u^2}}$$

Kumbhakar and Lovell (2000) describe use of this model and several variants.

4.2. Approximation by Hermite Quadrature

Butler and Moffitt's (1982) approach is based on models in which u_i has a normal distribution (though, in principle, it could be extended to others). If u_i is normally distributed with zero mean - the assumption is innocent in index function models as long as there is a constant term - then,

$$\int_{u_i} \left(\prod_{t=1}^{T(i)} g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right) h(u_i \mid \boldsymbol{\theta})\, du_i$$

$$= \frac{1}{\sigma_u \sqrt{2\pi}} \int_{-\infty}^{\infty} \prod_{t=1}^{T(i)} g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + u_i, \boldsymbol{\theta}) \exp\left(\frac{-u_i^2}{2\sigma^2} \right) du_i .$$

By a suitable change of variable for u_i, the integral can be written in the form

$$F = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \prod_{t=1}^{T(i)} g(y_{it}, \beta'\mathbf{x}_{it} + \tau v_i, \theta) \exp\left(-v_i^2\right) dv_i .$$

The function is in a form can be approximated very accurately with Gauss-Hermite quadrature, which eliminates the integration. Thus, the log likelihood function can be approximated with

$$\log L_h = \sum_{i=1}^{N} \log\left(\sum_{h=1}^{H} w_h \prod_{t=1}^{T(i)} g(y_{it}, \beta'\mathbf{x}_{it} + \tau z_h \mid \theta) \right)$$

where w_h and z_h are the weights and nodes for the Hermite quadrature of degree H. The log likelihood is fairly complicated, but can be maximized by conventional methods. This approach has been applied by numerous authors in the probit random effects context [Butler and Moffitt (1982), Heckman and Willis (1975), Guilkey and Murphy (1993) and many others] and in the tobit model [Greene (2000)]. In principle, the method could be applied in any model in which a normally distributed variable u_i appears, whether additive or not.[5] For example, Greene (2000) applies this technique in the Poisson model as an alternative to the more familiar log-gamma heterogeneity. The sample selection model is extended to the Poisson model in Greene (1994). One shortcoming of the approach is that it is difficult to apply to higher dimensional problems. Zabel (1992) and Tijman and Verbeek (1992) describe a bivariate application in the sample selection model, but extension of the quadrature approach beyond two dimensions appears to be impractical. [See also Bhat (1999).]

4.3. Simulated Maximum Likelihood

A third approach to the integration which has been used with great success in a rapidly growing literature is simulation. We observe, first, that the integral is an expectation;

$$\int_{u_i} \left(\prod_{t=1}^{T(i)} g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right) h(u_i \mid \boldsymbol{\theta}) du_i = E[F(u_i|\boldsymbol{\theta})]$$

The expectation has been computed thus far by integration. By the law of large numbers, if $(u_{i1},u_{i2},...,u_{iR})$ is a sample of iid draws from $h(u_i|\boldsymbol{\theta})$ then

[5] The Butler and Moffitt (1982) approach using Hermite quadrature is not particularly complicated, and has been built into most contemporary software, including, e.g., the Gauss library of routines. Still, there remains some skepticism in some of the applied literature about using this kind of approximation. Consider, for example, in van Ophem's (2000, p. 504) discussion of an extension of the sample selection model into a Poisson regression framework where he states "A technical disadvantage of this method is the introduction of an additional integral that has to be evaluated numerically in most cases." While true, the level of the disadvantage is extremely minor.

$$\text{plim}\ \frac{1}{R}\sum_{r=1}^{R} F(u_{ir}\,|\,\boldsymbol{\theta}) = E[F(u_i|\boldsymbol{\theta})]$$

This operation can be done by simulation using a random number generator. The simulated integral may then be inserted in the log likelihood, and maximization of the parameters can proceed from there. The pertinent questions are whether the simulation approach provides a stable, smooth, accurate enough approximation to the integral to make it suitable for maximum likelihood estimation and whether the resulting estimator can claim the properties that would hold for the exactly integrated counterpart. The approach was suggested early by Lerman and Manski (1983), and explored in depth in McFadden and Ruud (1994) [see, esp., Geweke, Keane, and Runkle (GKR) (1994, 1997)]. Keane (1994) and Elrod and Keane (1992) apply the method to discrete choice models. [See, as well, Section 3.1 of McFadden and Train (2000).] Hajivasilliou and Ruud (1994) is an influential survey of the method.

4.3.1. Simulation Estimation in Econometrics

Gourieroux and Monfort (1996) provide the essential statistical background for the simulated maximum likelihood estimator. We assume that the original maximum likelihood estimator as posed with the intractable integral is otherwise regular - if computable, the MLE would have the familiar properties, consistency, asymptotic normality, asymptotic efficiency, and invariance to smooth transformation. Let let $\boldsymbol{\beta}$ denote the full vector of unknown parameters being estimated and let $\mathbf{b}_{ML}$ denote the maximum likelihood estimator of the full parameter vector shown above, and let $\mathbf{b}_{SML}$ denote the simulated maximum likelihood (SML) estimator. Gourieroux and Monfort establish that if $\sqrt{N}/R \rightarrow 0$ and R and $N \rightarrow \infty$, then $\sqrt{N}\,(\,\mathbf{b}_{SML} - \boldsymbol{\beta})$ has the same limiting normal distribution with zero mean as $\sqrt{N}\,(\,\mathbf{b}_{ML} - \boldsymbol{\beta})$. That is, under the assumptions, the simulated maximum likelihood estimator and the maximum likelihood estimator are equivalent. The requirement that the number of draws, R, increase faster than the number of observations, N, is important to their result. The authors note that as a consequence, for "fixed R" the SML estimator is inconsistent. Since R is a parameter set by the analyst, the precise meaning of "fixed R" in this context is a bit ambiguous. On the other hand, the requirement is easily met by tying R to the sample size, as in, for example, $R = N^{1+\delta}$ for some positive δ. There does remain a finite R bias in the estimator, which the authors obtain as approximately equal to $(1/R)$ times a vector which is a finite vector of constants (see p. 44). Generalities are difficult, but the authors suggest that when the MLE is relatively "precise," the bias is likely to be small. In Munkin and Trivedi's (2000) Monte Carlo study of the effect, in samples of 1000 and numbers of replications areound 200 or 300 - note that their R is insufficient to obtain the consistency result - the bias of the estimator appears to be trivial.

4.3.2. Quasi-Monte Carlo Methods: The Halton Sequence

The results thus far are based on random sampling from the underlying distribution of $\mathbf{u}$. But, it is widely understood that the simulation method, itself, contributes to the variation of SML estimator. [See, e.g., Geweke (1995).] Authors have also found that judicious choice of the random draws for the simulation can be helpful in speeding up the convergence of this very computation intensive estimator. [See Bhat (1999).] One technique commonly used is antithetic sampling. [See Geweke (1995, 1998) and Ripley (1987).] The technique used involves not sampling R independent draws, but $R/2$ independent pairs of draws where the members of the pair are negatively correlated. One technique often used, for example is to pair each draw $\mathbf{u}_{ir}$ with

$-\mathbf{u}_{ir}$. (A loose end in the discussion at this point concerns what becomes of the finite simulation bias in the estimator. The results in Gourieroux and Monfort hinge on random sampling.)

Quasi Monte Carlo (QMC) methods are based on an integration technique that replaces the pseudo random draws of the Monte Carlo integration with a grid of "cleverly" selected points which are nonrandom but provide more uniform coverage of the domain of the integral. The logic of the technique is that randomness of the draws used in the integral is not the objective in the calculation. Convergence of the average to the expectation (integral) is, and this can be achieved by other types of sequences. A number of such strategies is surveyed in Bhat (1999), Sloan and Wozniakowski (1998) and Morokoff and Caflisch (1995). The advantage of QMC as opposed to MC integration is that for some types of sequences, the accuracy is far greater, convergence is much faster and the simulation variance is smaller. For the one we will advocate here, Halton sequences, Bhat (1995) found relative efficiencies of the QMC method to the MC method on the order of ten or twenty to one.

Monte Carlo simulation based estimation uses a random number to produce the draws from a specified distribution. The central component of the approach is draws from the standard continuous uniform distribution, $U[0,1]$. Draws from other distributions are obtained from these by using the inverse probability transformation. In particular, if u_i is one draw from $U[0,1]$, then $v_i = \Phi^{-1}(u_i)$ produces a draw from the standard normal distribution; v_i can then be unstandardized by the further transformation $z_i = \sigma v_i + \mu$. Draws from other distributions used, e.g., in Train (1999) are the uniform [-1,1] distribution for which $v_i = 2u_i - 1$ and the tent distribution, for which $v_i = \sqrt{2u_i} - 1$ if $u_i \leq 0.5$, $v_i = 1 - \sqrt{2u_i - 1}$ otherwise. Geweke (1995), and Geweke, Hajivassiliou, and Keane (1994) discuss simulation from the multivariate truncated normal distribution with this method.

The Halton sequence is generated as follows: Let r be a prime number larger than 2. Expand the sequence of integers $g = 1,...$ in terms of the base r as

$$g = \sum_{i=0}^{I} b_i r^i \text{ where by construction, } 0 \leq b_i \leq r - 1 \text{ and } r^I \leq g < r^{I+1}.$$

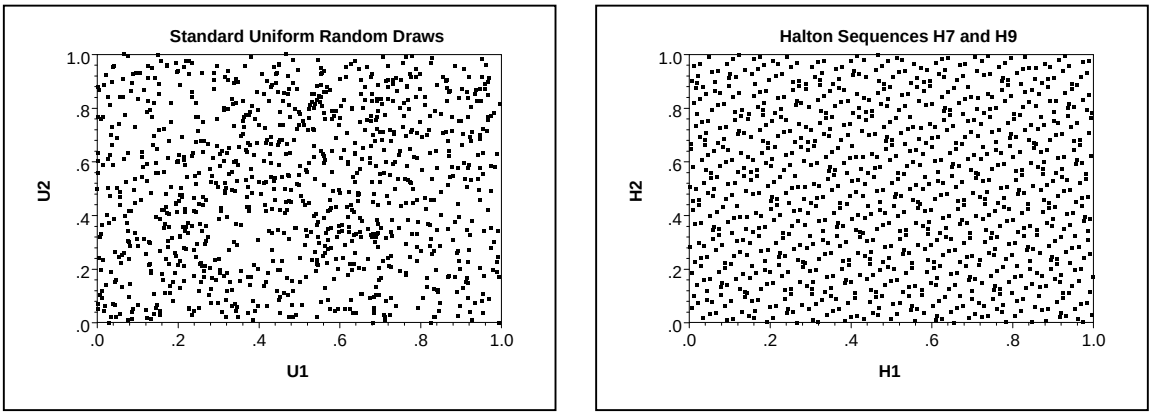
The Halton sequence of values that corresponds to this series is

$$H_r(g) = \sum_{i=0}^{I} b_i r^{-i-1}$$

For example, using base 5, the integer 37 has $b_0 = 2$, $b_1 = 2$, and $b_3 = 1$. Then

$$H_5(37) = 2\times5^{-1} + 2\times5^{-2} + 1\times5^{-3} = 0.448.$$

The sequence of Halton values is efficiently spread over the unit interval. The sequence is not random as the sequence of pseudo-random numbers is. The figures below show two sequences of Halton draws and two sequences of psuedorandom draws. The Halton draws are based on $r = 7$ and $r = 9$. The clumping evident in the figure on the left is the feature (among other others) that necessitates large samples for simulations.



4.3.3. Applications

The literature on discrete choice modeling now contains a surfeit of successful applications of this approach, notably the celebrated discrete choice analysis by Berry, Pakes, and Levinson (1995), and it is growing rapidly. Train (1998), Revelt and Train (1998) and McFadden and Train (2000) have documented at length an extension related to the one that we will develop here. Baltagi (1995) discusses a number of the 1990-1995 vintage applications. Keane et al., have also provided a number of applications.

Several authors have explored variants of the model above. Nearly all of the received applications have been for discrete choice models. McFadden's (1989), Bhat (1999, 2000), and Train's (1998) applications deal with a form of the multinomial logit model. Keane (1994) and GKR considered multinomial discrete choice for a multinomial probit model. In this study, Keane also considered the possibility of allowing the $T(i)$ observations on the same u_i to evolve as an AR or MA process, rather than to be a string of repetitions of the same draw. Other applications include Elrod and Keane (1992) and Keane (1994) in which various forms of the process generating u_i are explored - the results suggest that the underlying process generating u_i is not the major source of variation in the estimates when incorporating random effects in a model. It appears from the surrounding discussion [see, e.g., Baltagi (1995)] that the simulation approach offers great promise in extending qualitative and limited response models. (Once again, McFadden and Train (2000) discuss this in detail.) Our results with this formulation suggest that even this enthusiastic conclusion greatly understates the case. We will develop an extension of the random effects model that goes well beyond the variants considered even by Keane and Elrod. The approach provides a natural, flexible approach for a wide range of nonlinear models, and, as shown below, allows a rich formulation of latent heterogeneity in behavioral models.

4.4. A Random Coefficients Model

The random coefficients model has a long history in econometrics beginning with Rao (1973), Hildreth and Houck (1968) and Swamy (1972), but almost exclusively limited to the linear regression model. This has largely defined the thinking on the subject. As noted in Section 1.4, in the linear regression model, a random parameter vector with variation around a fixed mean produces a groupwise heteroscedastic regression that can, in principle, be fit by two step feasible generalized least squares. In an early application of this principle Akin, Guilkey, and Sickles (1979) extended this to the ordered probit model. To demonstrate, we consider the binomial probit model, which contains all the necessary features. The model is defined by

$$y_i = \mathbf{1}(\boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i > 0)$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i$$

where $\varepsilon_i \sim N[0,1]$ and $\mathbf{v}_i \sim N_k[\mathbf{0},\boldsymbol{\Sigma}]$. The reduced form of the model is

$$y_i = \mathbf{1}(\boldsymbol{\beta}'\mathbf{x}_i + \mathbf{v}_i'\mathbf{x}_i + \varepsilon_i > 0)$$

$$= \mathbf{1}(\boldsymbol{\beta}'\mathbf{x}_i + w_i > 0)$$

where $w_i \sim N[0, 1 + \mathbf{x}_i'\boldsymbol{\Sigma}\mathbf{x}_i]$. This is a heteroscedastic probit model [see, e.g, Greene (2000, Chapter 19)] which is directly estimable by conventional methods. The log likelihood function is

$$\log L_i = \sum_{i=1}^{N} \log \Phi\left((2y_i - 1)\frac{\boldsymbol{\beta}'\mathbf{x}_i}{1 + \mathbf{x}_i\boldsymbol{\Sigma}\mathbf{x}_i} \right)$$

The identified parameters in the model - the row and column in $\boldsymbol{\Sigma}$ corresponding to the constant term must be set to zero - are estimable by familiar methods. The authors extend this to the ordered probit model in the usual way [see McKelvey and Zavoina (1975)]. Sepanski (2000) revisited this model in a panel data setting, and added a considerable complication,

$$y_{it} = \mathbf{1}(\boldsymbol{\beta}_i'\mathbf{x}_{it} + \gamma y_{i,t-1} + \varepsilon_i > 0).$$

Even with the lagged dependent variable, the resulting estimator turns out to be similar to Guilkey et al's. The important element in both is that model estimation does not require integration of the heterogeneity out of the function. The heterogeneity merely lays on top of the already specified regression disturbance. The fact that the distributions mesh the way they do is rather akin to the choice of a conjugate prior in Bayesian analysis. [See Zellner (1971) for discussion.]

4.5. The Random Parameters Model

Most of the applications cited above save for those in the preceding section McFadden, and Train (2000) (and the studies they cite) and Train (1998) represent extensions of the simple additive random effects model to settings outside the linear model. Consider, instead, a *random parameters* formulation

$f(y_{it}|\mathbf{x}_{it},\mathbf{v}_i) = g(y_{it}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2i}, \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\theta})$

$\boldsymbol{\beta}_1 = K_1$ nonrandom parameters

$\boldsymbol{\beta}_{2i} = \boldsymbol{\beta}_2 + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{v}_i$

$= K_2$ random parameters with mean $\boldsymbol{\beta}_2 + \boldsymbol{\Delta}\mathbf{z}_i$ and variance $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'$

$\mathbf{v}_i$ = a random vector with zero mean vector and covariance matrix $\mathbf{I}$

$\boldsymbol{\Gamma}$ = a constant, lower triangular matrix

Δ = a constant $K_2 \times K_z$ parameter matrix

$\mathbf{z}_i$ = a set of K_z time invariant measurable effects.

The random parameters model embodies individual specific heterogeneity in the marginal responses (parameters) of the model. Note that this does not necessarily assume an index function formulation (though in practice, it usually will). The density is a function of the random parameters, the fixed parameters, and the data. The simple random effects models considered thus far are a narrow special case in which only the constant term in the model is random and $\boldsymbol{\Delta} = \mathbf{0}$. But, this is far more general. One of the major shortcomings of the random effects model is that the effects might be correlated with the included variables. (This is what has motivated the much less parsimonious fixed effects model in the first case.) The exact nature of that correlation has been discussed in the literature, see, e.g., Zabel (1992) who suggests that since the effect is time invariant, if there is correlation, it makes sense to model it in terms of the group means. The preceding allows that, as well as more general formulations in which $\mathbf{z}_i$ is drawn from outside $\mathbf{x}_{it}$. Revelt and Train (1999), Bhat (1999, 2000), McFadden and Train (2000), and others have found this model to be extremely flexible and useable in a wide range of applications in discrete choice modeling. The extension to other models is straightforward, and natural. (The statistical properties of the estimator are pursued in a lengthy literature that includes Train (1999, 2000), Bhat (1999), Lerman and Manski (1983) and McFadden et al. (1994).) Gourieroux and Montfort's smoothness condition on the parameters in the model is met throughout.

Irrespective of the statistical issues, the random parameters model addresses an important consideration in the panel data model. Among the earliest explorations of the issue of 'parameter heterogeneity' is Zellner (1962) where the possibly serious effects of inappropriate aggregation of regression models was analyzed. The natural question arises, if there is heterogeneity in the statistical relationship (linear or otherwise) why should it be confined to the constant term in the model? Certainly that is a convenient assumption, but one that should be difficult to justify on economic grounds. As discussed at length in Pesaran, Smith, and Im (1996), when panels are short and estimation machinery sparse, the assumption might have a compelling appeal. In more contemporary settings, neither is the case, so estimators that build on and extend the ones considered here seem more appropriate. If nothing else, the shifting constant model ought to be considered a maintained hypothesis.[6] The counterargument based on small T inconsistency makes the case an ambiguous one. Certainly for panels of $T(i)$=2 (as commonly analyzed in the literature on semiparametric estimation) the whole question is a moot point. But, in larger panels as often used in cross country, firm, or industry studies, the question is less clear cut. Pesaran et al. (1996) and El-Gamal and Grether (1999) discuss this issue at some length.

The simulated log likelihood for this formulation is

$$\log L_s = \sum_{i=1}^{N} \log\left[\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T(i)} g(y_{it}, \boldsymbol{\beta}_{ir}, \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\theta})\right]$$

where

$$\boldsymbol{\beta}_{ir} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{v}_{ir} \end{pmatrix}$$

and $\mathbf{v}_{ir}$ is a group specific, (repeated) set of draws from the specified distribution. We have found this model to work remarkably well in a large range of situations. (See Appendix B.) An extension which allows v_i to vary through time is an AR(1) model,

[6] Weighed against this argument at least for linear models is Zellner's (1969) result that if primary interest is in an 'unbiased' effect of strictly exogenous regressors, then pooling in a random parameters model will allow estimation of that effect. The argument loses its force, even in this narrow context, in the presence of lagged dependent variables or nonrandom heterogeneity.

$$v_{it,kr} = \rho_k v_{i,t-1,kr}$$

(where *i,t,k,r* index group, period, parameter, and replication, respectively). We note, once again, that this approach has appeared in the literature already (e.g., Berry, et al. (1995), Train et al. (1999) and the applications cited therein), but does not appear to have been extended beyond models for discrete choices. The modeling template represents a general extension of the random parameters model to other models, including probit, tobit, ordered probability, count data models, the stochastic frontier model, and others. A list of applications appears in Appendix B. Indeed, this is a point at which the understanding based on the linear model is a bit misdirecting. Conventional use of the random parameters model is built around the Swamy (1971) formulation of the linear model, which necessitates not only a panel, but one which is deep enough to allow each group to provide its own set of estimates, to be mixed in a generalized least squares procedure. [See also Swamy et al., (1988a,b and 1989).] But, *nothing in the preceding mandates panel data; the random parameters approach is amenable to cross section modeling as well*, and provides a general way to model individual specific heterogeneity. (This may seem counterintuitive, but consider that the familiar literature already contains applications of this, in certain duration models with latent heterogeneity (Weibull/gamma) and in the derivation of the negative binomial model from the Poisson regression.) We have applied this approach in a sample selection model for the Poison regression [Greene (1994)]. We note, as well, this modeling approach bears some similarity to the recently proposed GEE estimator. We will return to this in detail below.

4.6. Refinements

The preceding includes formulation of the random effects estimator proposed by Chamberlain (1980, 1984), where it is suggested that a useful formulation (using our notation) would be

$$u_i = \boldsymbol{\Delta}_i' \mathbf{z}_i + \varepsilon_i.$$

In the model with only a random constant term, this is exactly the model suggested above, where the set of coefficients is the single row in $\boldsymbol{\Delta}$ and ε_i would be $\boldsymbol{\Gamma}_{11} v_i$.

Second, this model would allow formulation of multiple equations of a SUR type. Through a nondiagonal $\boldsymbol{\Gamma}$, the model allows correlation across the parameters. Consider a two period panel where, rather than representing different periods, "period 1" is the first equation and "period 2" is the second. By allowing each equation to have its own random constant, and allowing these constants to be correlated, we obtain a two equation seemingly unrelated equations model - note that these are not linear regressions. In principle, this can be extended to more than two equations. (Full simultaneity and different types of equations would be useful extensions remaining to be derived.) This method of extending models to multiple equations in a nonlinear framework would differ somewhat from other approaches often suggested. Essentially, the correlation between two variates is induced by correlation of the two conditional means. Consider, for example, the Poisson regression model. One approach to modeling a bivariate Poisson model is to specify three independent Poisson models, w_1, w_2, and z. We then obtain a bivariate Poisson by specifying $y_1 = w_1 + z$ and $y_2 = w_2 + z$. The problem with this approach is that it forces the correlation between the two variables to be positive. It is not hard to construct applications in which exactly the opposite would be expected. [See, e.g., Gurmu and Elder (1998) who study demand for health care services, where frequency of illness is negatively correlated frequency of preventive measures. With a simple random effects approach, the

preceding is equivalent to modeling $E[y_{ij}|\mathbf{x}_{ij}] = \exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij})$ where $\text{cov}[\varepsilon_{i1},\varepsilon_{i2}] = \rho_{12}$. This is not a bivariate Poisson model as such. Whether this is actually a reasonable way to model joint determination of two Poisson variates remains to be investigated. (A related approach based on embedding bivariate heterogeneity in a model is investigated by Lee (1983) and van Ophem (1999, 2000).

The conditional means approach is, in fact, the approach taken by Munkin and Trivedi (1999), though with a slightly different strategy.[7] They begin with two Poisson distributed random variables, y_j each with its own displaced mean, $E[y_j|v_j] = \exp(\boldsymbol{\beta}_j'\mathbf{x}_j + v_j)$, $j = 1,2$. In their formulation, (v_1,v_2) have a bivariate normal distribution with zero means, standard deviations σ_1, σ_2, and correlation ρ. Their approach is a direct attack on the full likelihood function,

$$\log L = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{Poisson}[y_1 \mid \lambda_1(v_1)]\text{Poisson}[y_2 \mid \lambda_2(v_2)]\phi_2(v_1, v_2 \mid 0,0,\sigma_1,\sigma_2,\rho)\, dv_1 dv_2$$

where $\phi_2(.)$ denotes the density of the bivariate normal distribution. A major difficulty arises in evaluating this integral, so they turn to simulation, instead. The authors ultimately use a fairly complicated simulation approach involving a sampling importance function [see Greene (2000), Chapter 5] and a transformation of the original problem, but for our purposes, is is more useful to examine the method they first considered, then dismissed. Since the integral cannot be computed, but can be simulated, an alternative approach is to maximize the simulated log likelihood. They begin by noting that the two correlated random variables are generated from two standard normal primitive draws, ε_1 and ε_2, by $v_1 = \sigma_1\varepsilon_1$ and $v_2 = \sigma_2\left(\rho v_1 + (1-\rho^2)^{1/2} v_2\right)$. The simulated log likelihood is

$$\log L_s = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \text{Poisson}[y_{1i} \mid \lambda_{1i}(v_{1ir})]\text{Poisson}[y_{2i} \mid \lambda_{2i}(v_{2ir})]$$

They then substitute the expressions for v_1 and v_2, and maximize with respect to β_1, β_2, σ_1, σ_2, and ρ. The process becomes unstable as ρ approaches 1, which, apparently, characterizes their data. The random parameters approach suggested here would simplify the process. The same likelihood could be written in terms of $\sigma_1\varepsilon_{1ir}$ in the first density and $\sigma_2\varepsilon_{2ir} + \gamma_{21}\varepsilon_{1ir}$ in the second equation. The constraint on ρ becomes irrelevant, and γ_{21} becomes a free parameter. The desired underlying correlation, $\gamma_{21}/[\sigma_1(\sigma_2^2 + \gamma_{21}^2)^{1/2}]$ is computed ex post. This can be formulated in the model developed here by simply treating the two constant terms in the model as random correlated parameters.

4.7. Mechanics

The actual mechanics of estimation of the random parameters model are quite complex. Full details are provided in Greene (2001a, and 2001b). Appendix A provides a sketch of the procedure.

4.8. GEE Estimation

The preceding bears some resemblance to a recent development in the statistics literature, GEE (generalized equation estimation) modeling. [See Liang and Zeger (1986) and Diggle, Liang

[7] They also present a concise, useful survey of approaches to modeling bivariate counts. See, for example, Cameron and Johansson (1998).

and Zeger, (1994).] In point of fact, most of the internally consistent forms of GEE models (there are quite a few that are not) are contained in the random parameters model.

The GEE method of modeling panel data is an extension of Nelder and Wedderburn's (1972) and McCullagh and Nelder's (1983) Generalized Linear Models (GLIM) approach to specification. The generalized linear model is specified by a 'link' to the conditional mean function,

$$f(\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]) = \boldsymbol{\beta}'\mathbf{x}_{it},$$

and a 'family' of distributions,

$$y_{it} \mid \mathbf{x}_{it} \sim g(\boldsymbol{\beta}'\mathbf{x}_{it}, \boldsymbol{\theta})$$

where $\boldsymbol{\beta}$ and $\mathbf{x}_{it}$ are as already defined and $\boldsymbol{\theta}$ is zero or more ancillary parameters, such as the dispersion parameter in the negative binomial model (which is a GLIM). Many of the models mentioned earlier fit into this framework. The probit model has link function $f(.) = \Phi^{-1}(P)$ and Bernoulli distribution family; the classical normal linear regression has link function equal to the identity function and normal distribution family; and the Poisson regression model has a logarithmic link function and Poisson family. More generally, for any single index binary choice model, if $\mathrm{Prob}[y_{it} = 1] = F(\boldsymbol{\beta}'\mathbf{x}_{it})$, then this function is the conditional mean, and the link function is simply (by definition)

$$f(\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]) = F^{-1}[F(\boldsymbol{\beta}'\mathbf{x}_{it})] = \boldsymbol{\beta}'\mathbf{x}_{it}.$$

This captures many binary choice models, including probit, logit, Gompertz, complementary log log and Burr (scobit). A like result holds for the count models, Poisson and negative binomial, for which the link is simply the log function. So far, nothing has been added to models that are already widely familiar. The aforementioned authors demonstrate that the models which fit in this class can be fit by a kind of iterated weighted least squares, which is one of the reasons that GLIM modeling has gained such currency. (See below.) In the absence of a preprogrammed routine, it is easy to do.

One can create a vast array of models by crossing a menu of link functions with a second menu of distributional families. Consider, for example, the following matrix (which does not nearly exhaust all the possibilities). We choose four distributional families to provide models for the indicated commonly used kinds of random variables:

Random Variables		Link Functions				
Type of R.V.	**Family**	Identity	Logit	Probit	Logarithmic	Reciprocal
Binary	Bernoulli	X	•	•	X	X
Continuous	Normal	•	•	•	•	•
Count	Poisson	X	X	X	•	X
Nonnegative	Gamma	X	X	X	•	X

There is no theoretical restriction on the mesh between link and family. But, in fact, most of the combinations are internally inconsistent. For example, for the binary dependent variable, only the probit and logit links make sense; the others imply a conditional mean that is not bounded by zero and one. For the continuous random variable, any link could be chosen, but this just defines a linear or nonlinear regression model. For the count variable, only the log transformation insures an appropriate nonnegative mean. The logit and probit transformations imply a positive mean, but one would not want to formulate a model for counts that forces the conditional mean function to be a probability, so these make no sense either. The same considerations rule out all but the log transformation for the gamma family. The preceding lists most of the commonly used link

functions (some not listed are just alternative continuous distributions). More than half of our table is null, and of the nine combinations that work, five are just nonlinear regressions, which is a much broader class than this, and one would unduly restrict themselves if they limited themselves to the GLIM framework for nonlinear regression analysis. The upshot is that the GLIM framework adds little to what is already in place; in the end, GLIM is essentially an alternative (albeit, fairly efficient) method of estimating some models that are already routinely handled by most modern software with conventional maximum likelihood estimation.

GEE provides a variation of these already familiar models by extending them to panel data settings. The GEE approach adds a random effect to the GLIM for the panel of observations. The link function is redefined as

$$f(\mathrm{E}[y_{it} \mid \mathbf{x}_{it}]) = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},\ t = 1,...,T(i).$$

Now, consider some different approaches to formulating the $T(i)\times T(i)$ covariance matrix for the heterogeneity: Once again, we borrow some nomenclature from the GEE literature:

Independent:	$\mathrm{Corr}[\varepsilon_{it}, \varepsilon_{is}] = 0,\ t \neq s$
Exchangeable:	$\mathrm{Corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho,\ t \neq s$
AR(1):	$\mathrm{Corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho^{\lvert t-s\rvert},\ t \neq s$
Nonstationary:	$\mathrm{Corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho_{ts},\ t \neq s, \lvert t-s\rvert \leq g$
Unstructured:	$\mathrm{Corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho_{ts},\ t \neq s.$

The AR(1) model is precisely that used by Elrod and Keane (1992) and is the same as the random constants with AR(1) model discussed earlier. The exchangeable case is the now familiar random effects model. The GEE approach to estimation is a complex form of generalized method of moments in which the orthogonality conditions are induced by a series of approximations and assumptions about the form of the distribution (e.g. the method requires that the parametric family be of the linear exponential type). On the other hand, most of these models are already available in other forms. The first one is obvious - this is just the pooled estimator ignoring any group effects. The second is the random effects model. The differences between the most general form of the random parameters model and the GEE model are (1) received GEE estimators (e.g., Stata) include the latter two covariance structures while (2) the random parameters model allows random variation in parameters other than the constant term in the model. It is unclear which is more general. Keane et al. (1992) found some evidence that the form of the correlation structure in the latent effects makes little difference in the final estimates. If we restrict our attention to the AR(1) and exchangeable cases, then the random parameters model is far more flexible in that it does not require any assumptions about the form of the underlying density and it allows the heterogeneity to enter the model in more general forms. Finally, given a wide range of families crossed with link functions, the GEE estimator might well be applicable to a broader range of functional forms. However, reducing this set of models to those that do not imply an improper conditional mean or some other inappropriate restriction greatly reduces the size of this set. This dimension of the comparison appears to be uncertain. GEE has been widely used in the applied statistics literature, but appears to have made little appearance in econometrics. An exception is Brannas and Johansson (1995) which, as formulated, is a natural candidate. The write the Poisson regression model for observations t=1,...,$T(i)$ in terms of the structural regression function and a $T(i)$ vector of multiplicative disturbances with unrestricted covariance matrix; $\mathrm{E}[y_{it}|\varepsilon_{it}] = \varepsilon_{it}\exp(\boldsymbol{\beta}'\mathbf{x}_{it})$.

4.9. A Bayesian Approach

Though our focus here has been on classical estimation, this is a convenient point to note a contribution in the literature on Bayesian estimation. As demonstrated above, the random effects probit model represents one of the simpler applications of all of the random parameters model. It also provides a convenient vehicle for demonstration of the Gibbs sampler, Markov Chain Monte Carlo methods and data augmentation, as presented by Albert and Chib (1993) and Chib (1996) in their development of this particular model. Albert and Chib's estimator involves augmenting the data set with observations $z_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \mathbf{v}_i'\mathbf{x}_{it} + u_i$ - this has the form of the random parameters model with a random effect. The Gibbs sampler is based on cycling between sampling from the distributions of $\{z_{it}|\boldsymbol{\beta},\text{data}\}$ and $\{\boldsymbol{\beta}|z_{it},\text{data}\}$. The Gibbs sampler generates a sample from the marginal posterior distribution of $\boldsymbol{\beta}|$data. The estimates of the posterior mean and variance are then estimated using sample estimates of their population counterparts. There is a loose end in the model in the ultimate source of the prior for the underlying heterogeneity. On the other hand, once past that, Albert and Chib's method should generalize to settings other than the binomial probit model. As they note, the method obviates computation of the likelihood (and, by extension, its derivatives). Once the Gibbs sampler has been executed, estimates of the parameters of the posterior distribution are computed by simple sums of sample observations. Chib, Greenberg, and Winkelmann (1998) have applied the same technique to the Poisson regression model with random coefficients.

5. Latent Class Models

To define the *latent class model*, it is useful to write the random parameters formulation as

$$f(y_{it}|\mathbf{x}_{it},\boldsymbol{\beta}_i,\boldsymbol{\theta}) = g(y_{it}, \boldsymbol{\beta}_i, \mathbf{x}_{it}, \boldsymbol{\theta})$$

$$f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\beta}, \Delta, \boldsymbol{\theta}) = h(\mathbf{v}_i - \boldsymbol{\beta} - \Delta\mathbf{z}_i).$$

The unconditional density is

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\beta},\Delta,\boldsymbol{\theta}) = E_{\beta i}[g(y_{it}, \boldsymbol{\beta}_i, \mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\theta})]$$

$$= \int_{\mathbf{v}_i} g(y_{it}, \boldsymbol{\beta}_i, \mathbf{x}_{it}, \boldsymbol{\theta})\, h(\mathbf{v}_i - \boldsymbol{\beta} - \Delta\mathbf{z}_i)\, d\mathbf{v}_i$$

which is what we have analyzed in the preceding section.

The density of $\boldsymbol{\beta}_i$ is the mixing distribution. The preceding has assumed this is a continuous distribution. Suppose the mixture distribution has finite, discrete support. The resulting formulation is,

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\beta}, \Delta, \boldsymbol{\theta}) = \sum_{j=1}^{M} p_j\, g(y_{it}, \boldsymbol{\beta}_j(\mathbf{z}_i, \boldsymbol{\beta}, \Delta), \mathbf{x}_{it}, \boldsymbol{\theta})$$

where it remains to parameterize the regime probabilities, p_j and the structural parameters in the regime, $\boldsymbol{\beta}_j$. Note that we continue to carry a common set of parameters, $\boldsymbol{\theta}$. Estimation is over the regime probabilities themselves, the lower level parameters assumed to generate $\boldsymbol{\beta}_j$ and the ancillary parameters, $\boldsymbol{\theta}$.

5.1. Applications of Regime Switching Models

Latent class models have appeared at many points in the econometrics literature and in many apparently different forms that can be construed as latent class models. Most of the development has been outside the context of panel data modeling, so we begin the review with what are essentially cross section applications.

Among the earliest applications in econometrics is the mixture of normals problem explored by Quandt and Ramsey (1978),

$$f(y_i|\mu_j,\sigma_j^2) = \frac{1}{\sigma_j}\phi\left(\frac{y_i - \mu_j}{\sigma_j}\right), j = 1,2$$

$$p_1 = \lambda,\ 0 < \lambda < 1$$

$$p_2 = 1 - \lambda.$$

An extensive symposium on the subject produced, e.g., the method of moment generating functions estimator as an estimation strategy for the five unknown parameters.

In the switching regressions model [see Maddala and Nelson (1975)], the case in which there is no regime separation indicator can be cast as a direct extension of the model immediately above, in which

$$\mu_{ij} = \boldsymbol{\beta}_j'\mathbf{x}_i,\ j = 1,2.$$

but all other aspects are the same. Poirier and Ruud (1981) develop some of the underlying theory of the model, including the difficulty in estimation. The fundamental identification issue is discussed at length in Maddala (1983) and Goldfeld and Quandt (1975). Kiefer (1979, 1980a, 1980b) provides detailed analysis of the difficulty of estimating this model - for some parameter values, the likelihood function is unbounded.

Keane and Geweke (1999) have extended a latent class model to the probit setting;

$$y_i = \mathbf{1}(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i > 0)$$

$$f(\varepsilon_i) = \sum_{j=1}^{M} p_j \frac{1}{\sigma_j}\phi\left(\frac{\varepsilon_i - \mu_j}{\sigma_j}\right)$$

so the resulting probit model would be

$$\text{Prob}[y_i = 1] = \int_{-\beta'\mathbf{x}_i}^{\infty} \left(\sum_{j=1}^{M} p_j \left(\frac{1}{\sigma_j}\right)\Phi\left(\frac{\varepsilon_i - \mu_j}{\sigma_j}\right)\right) d\varepsilon_i$$

This is not quite the same as the formulation above, and is actually considerably more complex to estimate. The authors pursue a Bayesian approach to estimation, following results in Geweke (1993), Gelfand and Dey (1994) and Geweke (1997).

We note, finally, some related applications. The idea of a latent regime switch has been carried into the density, rather than the parameters, in the zero inflated Poisson models [see Lambert (1992) and Greene (1993)] and, of course, in a vast literature on policy regime switching in macroeconomics. In the latter, the time series nature of the applications makes them qualitatively different from those of interest here. The usual approach in macroeconometrics has involved Markov chain switching models and Kalman filter mechanisms. [See Hamilton (1995).] Finally, researchers in marketing have long used techniques such as conjoint analysis to identify latent patterns in consumption data. More recently, a considerable amount of formal analysis has been done using latent class structures, as in Vermunt and Magodson (1999a, 1999b, 2000) and Hagenar and McCutcheon (2001, forthcoming). (Note that these applications are typically not in the regression style that characterizes most of the econometric applications listed thus far.)

5.2. A Simple Latent Class Model

For present purposes, it is useful to view the structure as one of a discrete distribution of latent heterogeneity. Heckman and Singer (1984) suggested this approach as a semiparametric model for heterogeneity in a duration model.[8] In the formulation shown earlier, this would be

$$f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{j=1}^{M} p_j\, g(y_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_j, \boldsymbol{\theta}),\ 0 \le p_j \le 1, \Sigma_j p_j = 1.$$

(See also Laird (1978) Estimation of this model by direct maximization of the log likelihood is not especially difficult. The class probabilities must be constrained to sum to 1. We have found that a simple approach is to reparameterize them as a set of logit probabilities,

[8] They noted that the discrete approach might prove useful even if the heterogeneity were continuously distributed. One of their arguments was that formal modeling of a continuous latent random variable might produce an overparameterized model and degrade the estimation of the parameters of interest.

$$p_j = \frac{e^{\theta_j}}{\sum_{j=1}^{M} e^{\theta_j}}, j = 1,...,M, \theta_M = 0.$$

The resulting log likelihood is a continuous function of the parameters, and maximization is straightforward. We do find that the number of classes that can be identified is likely to be relatively small (on the order of five or less) and in general, (as might be expected), the less rich is the panel data set in terms of cross group variation, the more difficult it is to estimate this model. Note, as well, that the model is only weakly identified at the very best by a cross section.

Estimation produces values for the structural parameters, β, α_j and θ and the prior probabilities, p_j. For prediction purposes, one might be interested in the posterior class probabilities,

$$\begin{aligned}\text{Prob(class } j \mid \text{observation } i) &= \frac{f(\text{observation } i \mid \text{class } j)\,\text{Prob(class } j)}{\sum_{j=1}^{M} f(\text{observation } i \mid \text{class } j)\,\text{Prob(class } j)} \\ &= \frac{g\left(y_{it} \mid \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_j, \theta\right)p_j}{\sum_{j=1}^{M} g\left(y_{it} \mid \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_j, \theta\right)p_j} \\ &= w_{ij}.\end{aligned}$$

5.3. Extensions of the Latent Class Model

Two extensions of the latent class model seem natural. First, there is no reason to restrict the cross class variation to the constant term in the model. Thus, we rewrite the index function as

$$z_{it} = \boldsymbol{\beta}_j'\mathbf{x}_{it}$$

The previous model is now the special case in which only the constant term differs across classes. Second, we can parameterize the class probabilities, as in

$$p_{ij} = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{M} e^{\theta_{ij}}}, j = 1,...,M, \theta_M = 0, \ \theta_{ij} = \boldsymbol{\gamma}_j'\mathbf{z}_i.$$

In their analysis of criminal careers, Nagin and Land (1993) suggest both of these extensions, and go yet one step further in allowing $\mathbf{z}_i$ to evolve through time - in their study, the class probabilities are made dependent on the age of the individual.[9] Wang, Cockburn, and Puterman (1998) use a slightly less general form of the Nagin and Land model in a Monte Carlo study of the Patents and R&D relationship.

[9] Nagin and Land (1993) also extend the model to a mixture of distributions by layering the zero inflation model (intermittency model in their terms) on top of the latent class model. Thus, the structural form of their model involves a regime switch between active (R=1) and inactive (R=0) criminals and a regime for incidents that is modeled along the lines suggested above. See, also, Land, McCall, and Nagin (1994, 1995).

5.4. Applications of the Latent Class Model

As noted, the latent class formulation has provided an attractive platform for modeling latent heterogeneity. A number of applications have employed it under a variety of rubrics. The aforementioned applications, Wang et al. (1998) and Wedel et al. (1993) used the Poisson regression model to study counts of events. Deb and Trivedi used the same approach as Wedel et al., but they extended the modeling framework to the negative binomial model. Another counterpart to Wedel et al. is Brannas and Rosenqvist (1994), however their model was much less ambitious; they based the regression part on Heckman and Singer's argument for modeling latent heterogeneity, and shifted only the constant term. Tsionas (2000) has extended the stochastic frontier model in which the heterogeneity appears in the distribution of the one sided, or "efficiency" term in the compound disturbance. Phillips has applied the model to the linear regression model (1994) and to the variance components in a random effects linear model (2000). Wedel, DeSarbo, Bult and Ramaswamy (1993) have formally examined precisely the extended model suggested above, once again in the Poisson regression setting - they fit the class probabilities as simple scalars to be estimated. Nagin and Land, in contrast, fit the full model with heterogeneity in the latent class probabilities.

5.5. ML and EM Estimation of the Latent Class Model

Even with the full specification of heterogeneity in the class probabilities, the log likelihood function is not particularly complicated. A frontal assault on optimization will usually be successful. On the other hand, there is good reason to pursue other approaches to optimization.

Wedel et al. expend a fair amount of effort on imposing the adding up constraint on the prior probabilities. They ultimately accomplish this through a Lagrangean approach. As can be seen above, a simple reparameterization of the probabilities achieves the same end with much less effort. It is noteworthy that Nagin and Land (1993) and Wang et al. (1998) used this same parameterization. This does require a further constraint, that one of the parameters (or parameter vectors) be constrained to zero, but imposition of such a constraint is trivial. Brannas and Rosenqvist came up against the same difficulty. The probabilities in their model are forced to lie in the unit interval by using the parameterization $p_j = 1/[1+\exp(-\theta_j)]$ with θ_j unrestricted. This does solve the problem, but they did not impose the adding up constraint, $\Sigma_j p_j = 1$ in their model; they simply estimated the first $\theta_1,...,\theta_{M-1}$ without restriction and compued p_M residually, a procedure that is not guaranteed to succeed. The logit form of the structural latent class probabilities suggested by Nagin and Land (1993) is a simple and convenient approach that appears to have been used in about half of received applications.

A second issue concerns the estimation algorithm. Heckman and Singer (1984) advocated the EM algorithm for this model, reasoning that if the assumed *M* were larger than the true value, then the maximum of the likelihood located by the analyst would be on a ridge in the parameter space. [See Dempster, Laird, and Rubin (1977).] Thus, gradient methods would be inherently unstable. Since one would not know a priori the correct value of *M*, choice of the EM method of maximizing the likelihood function is not a solution to the problem. The EM method is an algorithm with (observed) very desirable stability properties. It also has the property that successive iterations always produces increases in the likelihood function. But, in the end, it remains a method of maximizing the same log likelihood as the ordinary gradient methods. The tradeoff is that the EM method is notoriously slow to converge in many cases. (See El-Gamal and Grether (1999), for example). Whether this really solves the problem is a strictly empirical question. On the other hand, the EM method for this problem turns out to be surprisingly easy.

To implement the EM method, Wedel et al. used the following approach: Let $u_{ij} = 1$ if individual i is a member of class j and zero otherwise. We treat u_{ij} as missing data to be estimated. The joint density of M u_{ij}s is multinomial with probabilities p_j;

$$f(u_1, u_2, ..., u_M) = \prod_{j=1}^{M} p_{ij}^{u_{ij}} .$$

The complete data log likelihood is, therefore

$$\log L_c = \sum_{i=1}^{N} \sum_{j=1}^{M} u_{ij} \log g(y_i \mid \mathbf{x}_i, \beta_j) + u_{ij} \log p_{ij}$$

The EM algorithm is used to maximize this log likelihood function. The expectation (E) step of the process involves obtaining the expectation of this log likelihood conditioned over the unobserved data. This ultimately involves replacing the u_{ij}s in log L_c with the posterior probabilities derived above (computed at the current estimates of the other parameters). The maximization (M) step then involves maximizing the resulting conditional log likelihood with these estimated posterior probabilities treated as known. Conditioned on the posteriors, E[logL_c] factors into two parts that may be maximized separately. Let w_{ij} denote the estimate posteriors. By construction, $\Sigma_j w_{ij} = 1$. The first part of the log likelihood becomes a weighted log likelihood with known weights for which the likelihood equations are

$$(*) \qquad \frac{\partial E[\log L_c]}{\partial \beta_j} = \sum_{i=1}^{N} w_{ij} \frac{\partial \log g(y_i \mid \mathbf{x}_i, \beta_j)}{\partial \beta_j} = \mathbf{0}.$$

This involves simply maximizing a weighted log likelihood for each class parameter vector. The maximum likelihood estimators of the class probabilities are just the sample averages of the estimated weights;

$$(**a) \qquad \hat{p}_j = \frac{\sum_{i=1}^{N} w_{ij}}{N}$$

If the logistic parameterization has been used, then the estimated θ_j is derived from $\theta_j = \log(p_j/p_M)$. Note also that if the Nagin and Land (1993) and Wang et al (1998) formulation of the prior probabilities as a function of sample data has been used, then the conditional log likelihood function at the M step for these parameters is a weighted multinomial logit log likelihood, which will require an iterative solution. That is, if

$$p_{ij} = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{M} e^{\theta_{ij}}}, j = 1, ..., M, \theta_M = 0, \ \theta_{ij} = \boldsymbol{\gamma}_j' \mathbf{z}_i.$$

then the transient solution for the structural parameters, γ_j is the solutions to the likelihood equations

$$(**b) \qquad \frac{\partial E[\log L_c]}{\partial \boldsymbol{\gamma}_j} = \sum_{i=1}^{N} (w_{ij} - \hat{p}_{ij}) \mathbf{z}_i = \mathbf{0}.$$

This is precisely the first order conditions for the multinomial logit model with proportion rather than individual data for the dependent variable (the weights).

The steps in the EM method are

(1) Obtain initial values of β_j, θ and, if part of the model, γ_j. The results of a single class model, replicated for all classes are one possibility.
(2) Compute prior probabilities. Compute weights (posterior class probabilities) w_{ij} based on current estimates of parameters.
(3) Solve (*) and (**) for new estimates of the parameters and the class probabilities.
(4) Assess whether parameter estimates have converged. If not, return to step (2)

5.5. EC Estimation

El-Gamal and Grether (1995, 1996) have proposed what they label the estimation-classification (EC) estimator. The model is essentially identical to the latent class estimator developed above with one crucial exception. The underlying theory involves a rather different interpretation. In their framework, the generation of observations from the latent classes is exact. That is, the data are generated by the process

$$g(y_{i1}, y_{i2},..., y_{i,T(i)} \mid \mathbf{X}_i, \beta_1, \beta_2,..., \beta_M) = \prod_{j=1}^{M} \left(g(y_{i1},... \mid \mathbf{X}_i, \beta_j)\right)^{\delta_{ij}}$$

where $\delta_{ij} \in \{0,1\}$ and $\sum_{j=1}^{M} \delta_{ij} = 1$. Thus, an observation (joint) is assumed to be a member of a specific class.[10] They note the resemblance of this to a fixed effects model. Citing Heckman and MaCurdy (1980) the authors consider the small T issue in this context, but argue that the consideration is misdirected. Although the classification indicators, δ_{ij} are estimated as parameters, the question of inconsistency is not an issue here. The small sample problem in this application concerns whether $T(i)$ is large enough to allow accurate classification.

Two major differences distinguish this estimation framework from the ones considered earlier. First, in a sample of N groups, there are $M^N/M!$ possible classifications, and, in principle, the estimator must search all of them to find the optimal one. For a large panel - recall, our interest in this survey has been in methods that could be applied when N = 50,000 - that would seem to make it impractical in the extreme. However, the authors present an algorithm that they argue makes this global search unnecessary. Second, the 'penalized' likelihood that they propose which must be maximized over the M slope vectors includes estimation of M as well. Thus, they simultaneously estimate the number of classes then the assignment of observations to these classes.

[10] Note the resemblance to discriminant analysis, which is (as usually analyzed) a special case of this model with M=2 and g(.) the joint normal distribution.

6. Conclusions

The applied econometrics literature has tended to view the selection of a random or fixed effects model as a Hobson's choice. The undesirable characteristics of the fixed effects model, notably the computational difficulties and, primarily, the inconsistency caused by the small $T(i)$ problem have rendered it a virtual neglected stepchild. As seen here, the practical issues may well be moot. The methodological problems remain. However, the pessimism suggested by examples which are doomed from the start - e.g., panel models with no regressors of substance and two periods, is surely overstated. There are many applications in which the group sizes are in the dozens or more, particularly in finance and in the long series derived from the PSID. In such cases, there might be room for more optimism. The point is that there is a compelling virtue of the fixed effects model as compared to the alternative, the random effects model. The assumption of zero correlation between latent heterogeneity and included, observed characteristics seems particularly severe. However, the problem can be made moot through extension of the random effects model to the random parameters model with unit specific means for the parameters.

Some analysts have suggested what might be viewed as a middle ground. A variety of non- and semiparametric estimators have been suggested. A notable exchange on the subject is Angrist et al. (2001) wherein it is argued that certain fairly ad hoc "approximations" provide the needed machinery. The commentary on Angrist's suggestions are fairly technical. Moffitt (2001), however, takes particular issue with the whole approach, arguing that substituting a demonstrably inconsistent model - in most cases the linear probability model for a well specified binary choice model - represents no solution at all. Suffice to say, the issue remains open for discussion.

The recent literature has suggested perhaps jumping between the horns of this dilemma through non- and semiparametric approaches. We would submit that this approach may be yet less informative than before. Consider, for example, Honore and Kyriazidou (2000) and. Kyriazidou (1997) as examples. In the context of "selection models" they show how one can tease out estimates of *structural* parameters of the model with minimal assumptions. The problem here is that the so called structural parameters in these models are essentially uninformative. They are not slopes of conditional means so they do not necessarily help in understanding behavior. The conditional means are not specified in these models, so neither are the estimated "parameters" helpful for prediction. At the risk of swimming against the incoming tide, it seems appropriate to ask whether the benefits to such weakly specified models are sufficient to outweigh the cost of rendering the estimates silent on questions that ultimately interest empirical researchers.

This paper has documented some extensions to a body of techniques that has existed in bits and pieces in the econometrics literature for some time. The end result is a collection of estimators that should extend the set of tools available to applied researchers. We acknowledge that the results apply to a narrow set, the minimal platform in fact, for specification of nonlinear panel data models. But, these results can certainly be extended. See, for example, Gourieroux and Monfort (1996) for some suggested applications. The recent literature also contains a host of applications to dynamic models that have extended these results in many directions. For static models, the contribution of Vella and Verbeek (1999) to nonlinear regression models with random effects also seems especially useful. Likewise, Woolridge (1995) offers some useful commentary for more general assumptions than made here.

References

Ahn, S. and P. Schmidt, "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 1995, pp. 3-38.

Akin, J., D. Guilkey and R. Sickles, A Random Coefficient Probit Model with an Application to a Study of Migration," *Journal of Econometrics*, 11, 1979, pp. 233-246.

Albert, J. and S. Chib, "Bayesian Analysis of Binary and Polytomous data," *Journal of The American Statistical Association*, 88, 1993, pp. 669-679.

Angrist, J., "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors," *Journal of Business and Economic Statistics*, 19, 1, 2001, pp. 2-15.

Arellano, M. and S. Bond, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 1991, pp. 277-297.

Arellano, M. and O. Bover, "Another Look at the Instrumental Variable Estimation of Error-Components Models," *Journal of Econometrics*, 68, 1995, pp. 29-51.

Baltagi, B., *Econometric Analysis of Panel Data*, John Wiley and Sons, New York, 1995.

Berry, S., J. Levinsohn and A. Pakes, "Automobile Prices in market Equilibrium," *Econometrica*, 63, 4, 1995, pp. 841-890.

Bhat, C., "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," Manuscript, Department of Civil Engineering, University of Texas, Austin, 1999.

Bhat, C. and S. Castelar, "A Unified Mixed Logit Framework for Modeling Revealed and Stated Preferences: Formulation and Application to Congestion Pricing Analysis in the San Francisco Bay Area," Manuscript, Department of Civil Engineering, University of Texas, Austin, 2000.

Bhat, C., "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," Manuscript, Department of Civil Engineering, University of Texas, Austin, 1999.

Brannas, K. and P. Johansson, "Panel Data Regressions for Counts," Manuscript, Department of Economics, University of Umea, Sweden, 1995.

Brannas, K. and G. Rosenqvist, "Semiparametric Estimation of Heterogeneous Count Data Models," *European Journal of Operational Research*, 76, 1994, pp. 247-258.

Brownstone, D. and K. Train, "Forecasting New Product Penetration with Flexible Substitution Patterns," *Journal of Econometrics*, 89, 1999, pp. 109-129.

Butler, J. and R. Moffitt, "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model," *Econometrica*, 50, 1982, pp. 761-764.

Cameron, A. and P. Johansson, "Bivariate Count Data Regression Using Series Expansions: With Applications," Discussion Paper, Department of Economics, University of California, Daavis, 1998.

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

Chib, S., E. Greenberg and R. Winkelmann, "Posterior Simulation and Bayes factor in Panel Count Data Models," Journal of Econometrics, 86, 1998, pp. 33-54.

Coelli, T., "A Guide to FRONTIER Version 4.1: A COmputer Program for Stochastic Frontier Production and Cost Estimation," Centre for Efficiency and Productivity Analysis, University of New England, Armidale, Australia, 1996.

Cornwell, C., P. Schmidt, and R. Sickles, "Production Frontiers with Cross Sectional and Time-Series Variation in Efficiency Levels," *Journal of Econometrics*, 46, 1990, pp. 185-200.

Crepon, B. and E. Duguet, "Research and Development, Competition and Innovation: Pseudo Maximum Likelihood and Simulated Maximum Likelihood Method Applied to Count Data Models with Heterogeneity," Journal of Econometrics, 79, 1997, pp. 355-378.

Deb, P. and P. Trivedi, "Demand for Medical Care by the Elderly: A Finite Mixture Approach," Journal of Applied Econometrics, 12, 3, 1997, pp. 313-336.

Dempster, A., N. Laird and D. Rubin, "Maximum Likelihood From Incomplete Data via the E.M. Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1, 1977, pp. 1-38.

Diggle, P., K. Liang and S. Zeger, *Analysis of Longitudinal Data*, Clarendon Press, Oxford, 1994.

El-Gamal, M. and D. Grether, "A Monte Carlo Study of EC Estimation in Panel Data Models with Limited Dependent Variables and Heterogeneity," in Hsiao, et al., eds., *Analysis of Panels and Limited Dependent Variable Models*, Cambridge University Press, Cambridge, 1999.

El-Gamal and D. Grether, "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of the American Statistical Association*, 90, 1995, pp. 1137-1145.

El-Gamal and D. Grether, "Unknown Heterogeneity, The EC-EM Algorithm, and Large T Approximation," SSRI Working Paper Number 9622, University of Wisconsin, Madison, 1996.

Elrod, T. and M. Keane, "A Factor Analytic Probit Model for Estimating Market Structure in Panel Data," *Journal of Marketing Research*, 1992.

Frisch, R., and F. Waugh, "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1, 1933, pp. 387-401.

Gelfand, A. and D. Dey, "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society, Series B*, 56, 1994, pp. 501-514.

Geweke, J., "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics*, 38, 1988, pp. 73-89.

Geweke, J., "Monte Carlo Simulation and Numerical Integration," Staff Research Report 192, Federal Reserve Bank of Minneapolis, 1995.

Geweke, J., "Posterior Simulators in Econometrics," in Kreps, D. and K. Wallis, eds., *Advances in Statistics and Econometrics: Theory and Applications, Vol III*, Cambridge University Press, Cambridge, 1997.

Geweke, J., M. Keane and D. Runkle, "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, 1994, pp. 609-632.

Geweke, J., M. Keane and D. Runkle, "Statistical Inference in the Multinomial Multiperiod Probit Model," Journal of *Econometrics*, 81, 1, 1997, pp. 125-166.

Goldfeld S. and R. Quandt, "Estimation in a Disequilibrium Model and the Value of Information," *Journal of Econometrics*, 3, 3, 1975, pp. 325-348.

Gourieroux, C. and A. Monfort, *Simulation Based Econometrics*, Oxford University Press, New York, 1996.

Greene, W., "Accounting for Excess Zeros and Sample Selection in the Poisson Regression Model," Working Paper Number 94-10, Department of Economics, Stern School of Business, NYU, 1994.

Greene, W., Econometric Analysis, 2nd ed., Macmillan, New York, 1993.

Greene, W., Econometric Analysis, 4th ed., Prentice Hall, Englewood Cliffs, 2000.

Greene, W., "Estimating a Random Parameters Model," manuscript, Department of Economics, Stern School of Business, NYU, 2001.

Greene, W., "Estimating Sample Selection Models with Panel Data," Manuscript, Department of Economics, Stern School of Business, NYU, 2001.

Greene, W., *LIMDEP*, Version 7.0, Econometric Software, Plainview, New York, 2000.

Guilkey, D., and J. Murphy, "Estimation and Testing in the Random Effects Probit Model," *Journal of Econometrics*, 59, 1993, pp. 301-317.

Gurmu, S., and J. Elder, "Estimation of Multivariate Count Regression Models With Applications to Health Care Utilization," Manuscript, Department of Economics, Gerogia State University, 1998.

Hagenars, J. and A. McCutcheon, *Advances in Latent Class Analysis*, Cambridge University Press, Cambridge, 2001 (forthcoming).

Hajivassiliou, V. and P. Ruud, "Classical Estimation Methods for LDV Models Using Simulation," In Engle, R. and D. McFadden, eds., Handbook of Econometrics, Vol. IV, North Holland, Amsterdam, 1994.

Hamilton, J., *Time Series Analysis*, Princeton University Press, Princeton, 1995.

Hausman, J., B. Hall and Z. Griliches, "Econometric Models for Count Data with an Application to the Patents - R&D Relationship," *Econometrica*, 52, 1984, pp. 909-938.

Hausman, J. and W. Taylor, "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1981, pp. 1377-1398.

Heckman, J. and B. Singer, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 1984, pp. 271-320.

Heckman, J. and MaCurdy, T., "A Life Cycle Model of Female Labor Supply," *Review of Economic Studies*, 47, 1980, pp. 247-283.

Heckman, J. and R. Willis, "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in Terlyckyi, N., ed., *Household Production and Consumption*, NBER, New York, 1975, pp. 99-138.

Heckman, J., "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in Manski, C. and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, 1981, pp. 114-178.

Hildreth, C. and J. Houck, "Some Estimators for a Linear Model with Random Coefficients," *Journal of the American Statistical Association*, 63, 1968, pp. 584-595.

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

Holtz-Eakin, D., "Testing for Individual Effects in Autoregressive Models," *Journal of Econometrics*, 39, 1988, pp. 297-307.

Holtz-Eakin, D., W. Newey and S. Rosen, "The Revenues-Expenditures Nexus: Evidence from Local Government Data," *International Economic Review*, 30, 1989, pp. 415-429.

Honore, B., "IV Estimation of Panel Data Tobit Models with Normal Errors," manuscript, Department of Economics, Princeton University, 1996.

Honore, B. and E. Kyriazidou, "Panel Data Discrete Choice Models with Lagged Dependent Variable Models," *Econometrica*, 68, 2000, pp. 839-874.

Honore, B., "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 1992, pp. 533-565.

Hsiao: C., *Analysis of Panel Data*, Cambridge University Press, Cambridge, 1993, pp. 159-164

Hsiao, C, "Logit and Probit Models," in Matyas, L. and Sevestre, P., eds., *The Econometrics of Panel Data: Handbook of Theory and Applications, Second Revised Edition*, Kluwer Academic Publishers, Dordrecht, 1996 pp. 410-447.

Judson, R. and A. Owen, "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists," *Economics Letters*, 65, 1999, pp. 9-15.

Keane, M., "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62, 1994, pp. 95-116.

Kiefer, N., "A Note on Regime Classification in Disequilibrium Models," *Review of Economic Studies*, 47, 1, 1980, pp. 637-639.

Kiefer, N., "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model," *Econometrica*, 46, 1978, pp. 427-434.

Kiefer, N., "On the Value of Sample Separation Information," *Econometrica*, 47, 1979, pp. 997-1003.

Krailo, M. and M. Pike, "Conditional Multivariate Logistic Analysis of Stratified Case-Control Studies," *Applied Statistics*, 44, 1, 1984, pp. 95-103.

Kyriazidou, E., "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1997, pp. 1335-1364.

Land, K., P. McCall, and D. Nagin, "Poisson and Mixed Poisson Regression Models: A Review of Applications, Including Recent Developments in Semiparametric Maximum Likelihood Methods," Manuscript, Department of Sociology, Duke University, 1994.

Land, K., P. McCall, and D. Nagin, "A Comparison of Poisson, Negative Binomial and Semiparametric Mixed Poisson Regression Modles with Empirical APplications to Criminal Careers Data," Manuscript, Department of Sociology, Duke University, 1995.

Laird, N., "Nonparametric Maximum Likelihood Estimation of Mixing Distributions," *Journal of the American Statistical Association*, 73, 1978, pp. 805-811.

Lambert, D., "Zero Inflated Poisson Regression, with an Application to Defects in Manufacturing," *Technometrics*, 34, 1, 1992, pp. 1-14.

Lechner, M. and M. Breitung, "Some GMM Estimation Methods and Specification Tests for Nonlinear Models," in Matyas, L. and Sevestre, P., *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Kluwer, Boston, 1996.

Lee, L., "Generalized Econometric Models with Selectivity," *Econometrica*, 51, 1983, pp. 507,512.

Lerman, S. and C. Manski, "On the Use of Simulated Frequencies to Approximate Choice Probabilities," In Manski, C and McFadden, D., eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, 1981

Liang, K. and S. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 1986, pp. 13-22.

Maddala, G., "Limited Dependent Variable Models Using Panel Data," *Journal of Human Resources*, 22, 3, 1987, pp. 307-338.

Maddala, G. and F. Nelson, "Specification Errors in Limited Dependent Variable Models," Working Paper number 96, National Bureau of Economic Research, Cambridge, 1975.

Manski, C., "Semiparametric Analysis of random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 1987, pp. 357-362.

Matyas, L. and Sevestre, P., eds., *The Econometrics of Panel Data: Handbook of Theory and Applications, Second Revised Edition*, Kluwer Academic Publishers, Dordrecht, 1996, pp. 410-447.

McCullagh, P. and J. Nelder, *Generalized Linear Models*, Chapman and Hall, New York, 1983.

McFadden, D., "Am Method of Simulated Moments for Estimation of Discrete Choice Models without Numerical Integration," *Econometrica*, 57, 1989, pp. 995-1026.

McFadden, D. and P. Ruud, "Estimation by Simulation," *Review of Economics and Statistics*, 76, 1994, pp. 591-608.

McFadden, D. and K. Train, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 2000, pp. 447-470.

Moffitt, R., "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Comment" *Journal of Business and Economic Statistics*, 19, 1, 2001, p 20.

Montalvo, J., "GMM Estimation of Count Panel Data Models with Fixed Effects and Predetermined Instruments," Journal of Business and Economic Statistics, 15, 1, 1997, pp. 82-89.

Morokoff, W., and R. Calflisch, "Quasi-Monte Carlo Integration," *Journal of Computational Physics*, 122, 1995, pp. 218-230.

Munkin, M. and P. Trivedi, "Econometric Analysis of a Self Selection Model with Multiple Outcomes Using Simulation-Based Estimation: An Application to the Demand for Healthcars," Manuscript, Department of Economics, Indiana University, 2000.

Nagin, D. and K. Land, "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model," *Criminology*, 31, 3, 1993, pp. 327-362.

Nelder, J. and R. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 1972, pp. 370-384.

Nijman, T. and M. Verbeek, "Nonresponse in Panel Data: The Impact on Estimates of Life Cycle Consumption Function," *Journal of Applied Econometrics*, 7, 1992, pp. 243-257.

Nerlove, M., "An Essay on the History of Panel Data Econometrics," Manuscript, Department of Agricultural and Resource Economics, University of Maryland, 2000.

Oberhofer, W. and J. Kmenta, "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models," *Econometrica*, 42, 1974, pp. 579-590.

Orme, C., "Two-Step Inference in Dynamic Non-Linear Panel Data Models," Manuscript, School of Economic Studies, University of Manchester, 1999.

Pesaran, H., R. Smith and K. Im, "Dynamic Linear Models for Heterogeneous Panels," in Matyas, L. and P. Sevestre, eds., *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Kluwer, Boston, 1996.

Philips, R., "Estimation of a Stratified Error Components Model," Manuscript, Department of Economics, George Washington University, 2000.

Philips, R., "Partially Adaptive Estimation via a Normal Mixture," *Journal of Econometrics*, 64, 1994, pp. 123-144.

Pitt, M. and L. Lee, "The Measurement and Sources of Technical Inefficiency in Indonesian Weaving Industry," *Journal of Development Economics*, 9, 1981, pp. 43-64.

Poirier, D. and P. Ruud, "On the Appropriateness of Endogenous Switching," *Journal of Econometrics*, 16, 2, 1981, pp. 249-256.

Quandt, R. and J. Ramsey, "Estimating Mixtures of normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 1978, pp. 730-738.

Rao, C., *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York, 1973.

Revelt, D. and K. Train, "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *Review of Economics and Statistics*, 80, 1998, pp. 1-11.

Ripley, B., *Stochastic Simulation*, John Wiley and Sons, New York, 1987.

Schmidt, P. and R. Sickles, "Production Frontiers and Panel Data," *Journal of Business and Economic Statistics*, 2, 1984, pp. 367-374.

Sepanski, J., "On a Random Coefficient Probit Model," *Communications in Statistics* - Theory and methods, 29, 11, 2000, pp. 2493-2505.

Sloan, J. and H. Wozniakowski, "When are Quasi-Monte Carlo Algorithms Efficient for High Dimensional Integrals," *Journal of Complexity*, 14, 1998, pp. 1-33.

Swamy, P., *Statistical Inference in Random Coefficient Regression Models*, Springer-Verlag, New York, 1971.

Swamy, P. and S. Arora, "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," *Econometrica*, 40, 1972, pp. 261-275.

Swamy, P., R. Conway, and M. LeBlanc, "The Stochastic Coefficients Approach to Econometric Modeling, Part I: A Critique of Fixed Coefficient Models," *The Journal of Agricultural Economic Research*, 40, 1988a, pp. 2-10.

Swamy, P., R. Conway, and M. LeBlanc, "The Stochastic Coefficients Approach to Econometric Modeling, Part II: Description and Motivation," *The Journal of Agricultural Economic Research*, 40, 1988b, pp. 21-30

Swamy, P., R. Conway, and M. LeBlanc, "The Stochastic Coefficients Approach to Econometric Modeling, Part III: Estimation, Stability Testing and Prediction," *The Journal of Agricultural Economic Research*, 41, 1989 pp. 4-20.

Train, K., "Recreation Demand Models with Taste Differences over People," *Land Economics*, 74, 1998, pp. 230-239.

Train, K., "Halton Sequences for Mixed Logit," Manuscript, Department of Economics, University of California, Berkeley, 1999.

Tsionas, E., "Non-normality in Stochastic Frontier Models: With an Application to U.S. Banking," *Journal of Productivity Analysis*, 2001, forthcoming.

van Ophem, H., "Modeling Selectivity in Count-Data Models," *Journal of Business and Economic Statistics*, 18, 4, 2000, pp. 503-511.

van Ophem, H., "A General Method to Estimated Correlated Discrete Random Variables," *Econometric Theory*, 15, 1999, pp. 228-237.

Vella, F. and M. Verbeek, "Two Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias," *Journal of Econometrics*, 90, 1999, pp. 239-263.

Verbeek, M., "On the Estimation of a Fixed Effects Model with Selectivity Bias," *Economics Letters*, 34, 1990, pp. 267-270.

Verbeek, M. and T. Nijman, "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 33, 3, 1992, pp. 681-703.

Vermunt, J. and J. Magidson, "Bi-Plots and Related Graphical Displays Based on Latent Class Factor and Cluster Models," Manuscript, Tilburg University, 1999b.

Vermunt, J. and J. Magidson, "Latent Class Cluster Analysis," Manuscript, Tilburg University, 1999a.

Vermunt, J. and J. Magidson, "Latent Class Models," Manuscript, Tilburg University, 2000.

Wang, P., I. Cockburn, and M. Puterman, "Analysis of Patent Data - A Mixed Poisson Regression Model Approach," *Journal of Business and Economic Statistics*, 16, 1, 1998, pp. 27-41.

Wedel, M., W. DeSarbo, J. Bult, and V. Ramaswamy, "A Latent Class Poisson Regression Model for Heterogeneous Count Data," *Journal of Applied Econometrics*, 8, 1993, pp. 397-411.

Woolridge, J., "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68, 1995, pp. 115-132.

Zabel, J., "Estimating Fixed and Random Effects Models with Selectivity," *Economics Letters*, 40, 1992, pp. 269-272.

Zavoina, R. and W. McKelvey, "A Statistical Model for the Analysis of Ordinal Data," *Journal of Mathematical Sociology*, Summer, 1975, pp. 103-120.

Zellner, A., An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias," *Journal of the American Statistical Association*, 57, 1962, pp. 500-509.

Zellner, A., *An Introduction to Bayesian Inference In Econometrics*, John Wiley and Sons, New York, 1971.

Zellner, A., "On the Aggregation Problem: A New Approach to a Troublesome Problem," in Fox, K et al., eds., *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*, Springer Verlag, Heidelberg, 1969.

Appendix A. Computation of the Random Parameters Model

Two models are used for $\mathbf{v}_{it}$:

Random Effects: $\mathbf{v}_{it} = \mathbf{v}_i$ for all t. This is the usual random effects form.

Autocorrelated [AR(1)] $\mathbf{v}_{it} = \mathbf{R}\mathbf{v}_{i,t-1} + \mathbf{u}_{it}$ where $\mathbf{R}$ is a diagonal matrix of coefficient specific autocorrelation coefficients and $\mathbf{u}_{it}$ satisfies the earlier specification for $\mathbf{v}_{it}$.

The remainder of the specification is

$\mathbf{\Gamma}$ = lower triangular or diagonal matrix which produces the covariance matrix of the random parameters, $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Gamma}'$ in the random effects form and $\mathbf{\Omega} = \mathbf{\Gamma}(\mathbf{I}-\mathbf{R}^2)^{-1}\mathbf{\Gamma}'$ in the AR(1) model.

$\mathbf{x}_{2it}$ = variables multiplied by $\boldsymbol{\beta}_{2it}$

$\boldsymbol{\beta}_{it}$ = $[\boldsymbol{\beta}_1', \boldsymbol{\beta}_{2it}']'$

$\mathbf{x}_{it}$ = $[\mathbf{x}_{1it}', \mathbf{x}_{2it}']'$

a_{it} = $\boldsymbol{\beta}_{it}'\mathbf{x}_{it}$

(We confine attention to index function models, though others are possible.)

$P(y_i|\mathbf{x}_{it}, \mathbf{z}_i, \mathbf{v}_{it}) = g(y_{it}, a_{it}, \boldsymbol{\theta})$ = the conditional density for the observed response.

The model assumes that parameters are randomly distributed with possibly heterogeneous (across individuals) mean

$$\mathrm{E}[\boldsymbol{\beta}_{it}|\,\mathbf{z}_i] = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i,$$

and

$$\mathrm{Var}[\boldsymbol{\beta}_{it}|\,\mathbf{z}_i] = \mathbf{\Omega}.$$

By construction, then,

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{\Gamma}\mathbf{v}_{it}.$$

Note that in the AR(1) form, $\boldsymbol{\beta}_{it}$ varies across time as well as individuals. It is convenient to analyze the model in this fully general form at this point. One can easily accommodate nonrandom parameters just by placing rows of zeros in the appropriate places in $\mathbf{\Gamma}$. A hierarchical parameter structure is accommodated with nonzero rows in $\boldsymbol{\Delta}$ with or without stochastic terms induced by nonzero terms in $\mathbf{\Gamma}$.

The true log likelihood function is

$$\log L = \Sigma_i \log L_i$$

where $\log L_i$ is the contribution of the ith individual (group) to the total. Conditioned on $\mathbf{v}_i$, the joint density for the ith group is

$$f[y_{i1},\ldots,y_{iTi} \mid \mathbf{x}_{it,\ldots,}\, \mathbf{z}_i, \mathbf{v}_{it},\ t = 1,\ldots,\ T(i)] = \prod_{t=1}^{T_i} g(y_{it}, \boldsymbol{\beta}_{it}'\mathbf{x}_{it}).$$

Since $\mathbf{v}_{it}$ is unobserved, it is necessary to obtain the unconditional log likelihood by taking the expectation of this over the distribution of $\mathbf{v}_{it}$. Thus,

$$L_i \mid \mathbf{v}_{it},\ t=1,\ldots,\ T(i) \ = \ \prod_{t=1}^{T_i} g(y_{it}, \boldsymbol{\beta}_{it}'\mathbf{x}_{it}) .$$

and,

$$L_i \ = \ E_{\mathbf{v}it}\,[L_i \mid \mathbf{v}_{it},\ t=1,\ldots,T(i)] = \int_{\text{Range of } \mathbf{v}_{it}} g(\mathbf{v}_{it}, t=1,\ldots,T(i)) \prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{it}'\mathbf{x}_{it})\, d\mathbf{v}_{it}$$

(Note that this is a multivariate integral.) Then, finally,

$$\log L \quad = \quad \sum_{i=1}^{N} \log L_i$$

For convenience in what follows, let $\boldsymbol{\Theta}$ = the full vector of all parameters in the model. The likelihood function is maximized by solving the likelihood equations:

$$\frac{\partial \log L}{\partial \boldsymbol{\Theta}} \ = \ \sum_{i=1}^{N} \frac{\partial \log L_i}{\partial \boldsymbol{\Theta}} \ = \ \mathbf{0}.$$

and note that these derivatives will likewise involve integration.

The integration is done by Monte Carlo simulation;

$$E_{\mathbf{v}it}\,[L_i \mid \mathbf{v}_{it},\ t=1,\ldots,\ T(i)] \ \approx \ \frac{1}{R} \sum_{r=1}^{R} L_i \mid \mathbf{v}_{itr}, t=1,\ldots,T(i) .$$

where $\mathbf{v}_{itr}$ is a set of $T(i)$ K_2-variate random draws from the joint distribution of $\mathbf{v}_{it}$. (I.e., it is a draw of a $T_i \times K_2$ random matrix. In the case of no autocorrelation, there is only one K_2-variate draw, which is then the same in all periods, in the fashion of a random effects model.) See Brownstone and Train (1999), Train (1998), and Revelt and Train (1998) for discussion. (Ken Train has numerous other papers on this subject which may be perused at his website.) The approximation improves with increased R and with increases in N, though the simulation variance which decreases with increases in R does not decrease with N.

The K_2 elements of $\mathbf{v}_{itr}$ are drawn as follows: We begin with a K_2 random vector $\mathbf{w}_{itr}$ which is either K_2 independent draws from the standard uniform [0,1] distribution or K_2 Halton draws from the mth Halton sequence, where m is the mth prime number in the sequence of K_2 prime numbers beginning with 2. [See Greene (2001a)]. The Halton values are also distributed in the unit interval. This primitive draw is then transformed to one of the following distributions, depending on the appropriate model. Train (1999, 2000) has suggested three possibilities:

Uniform[-1,1]: $u_{k,itr} = 2w_{k,itr} - 1$

Tent [-1,1] $u_{k,itr} = \mathbf{1}(w_{k,itr} \leq .5)[\sqrt{2w_{k,itr}} - 1] +$

$\mathbf{1}(w_{k,itr} > .5)[1 - \sqrt{2(1-w_{k,itr})}\ ]$

Normal[0,1] $u_{k,itr} = \Phi^{-1}(w_{k,itr})$

This produces a K_2 vector, $\mathbf{u}_{itr}$. Finally, $\mathbf{v}_{itr}$ is obtained as follows:

(1) No autocorrelation: $\mathbf{v}_{itr} = \mathbf{u}_{itr}$ for all t.

In this case, $\mathbf{w}_{itr}$ is drawn once for the entire set of $T(i)$ periods, and reused. This is the standard 'random effect' arrangement, in which the effect is the same in every period. In this case,

$\mathbf{w}_{itr} = \mathbf{w}_{ir}$, $\mathbf{u}_{itr} = \mathbf{u}_{ir}$, and $\mathbf{v}_{itr} = \mathbf{v}_{ir}$,

(2) AR1 model (autocorrelation): $v_{k,i1r} = [1/(1 - \rho_k^2)]\, u_{k,i1r}$

$v_{k,itr} = \rho_k\, v_{k,i,t-1,r} + u_{k,itr}$

This is the standard first order autocorrelation treatment, with the Prais-Winsten treatment for the first observation - this so as not to lose any observations due to differencing.

In the preceding derivation, it is stated that $\mathbf{\Omega} = \mathbf{\Gamma\Gamma'}$ is the covariance matrix of $\mathbf{\Gamma v}_{itr}$. This is true for the standard normal case. For the other two cases, a further scaling is needed. The variance of the uniform [-1,1] is the squared width over 12, or 1/3, so its standard deviation is $1/\sqrt{3}$ = .57735. The variance of the standardized tent distribution is 1/6. The standard deviation is therefore .40824.

With $\mathbf{v}_{itr}$ in hand, we form the rth draw on the random parameter, $\boldsymbol{\beta}_{itr}$ as follows:

$\boldsymbol{\beta}_{1itr}$ = $\boldsymbol{\beta}_1$ (K_1 nonrandom parameters - does not change with i, r, or t). This parameter vector is being estimated.

$\boldsymbol{\beta}_{2itr}$ = $\boldsymbol{\beta}_2 + \mathbf{\Delta z}_i + \mathbf{\Sigma v}_{itr} + \mathbf{\Pi v}_{itr}$ (K_2 random parameters)

= $\boldsymbol{\beta}_2 + \mathbf{\Delta z}_i + \mathbf{\Gamma v}_{itr}$ where $\mathbf{\Gamma} = \mathbf{\Sigma} + \mathbf{\Pi}$, $\mathbf{\Sigma}$ is diagonal and $\mathbf{\Pi}$ is the nonzero elements below the diagonal in $\mathbf{\Gamma}$.

The parameter vector, $\boldsymbol{\beta}_{itr}$ is now in hand.

The probability density function is formed by beginning with

$P_{itr} = g(y_{it}, \boldsymbol{\beta}_{itr}'\mathbf{x}_{it}, \boldsymbol{\theta})$

(Note, if this is the random effects model, then $\boldsymbol{\beta}_{itr}'\mathbf{x}_{it} = \boldsymbol{\beta}_{ir}'\mathbf{x}_{it}$.) The joint conditional probability for the ith individual is

$$P_{ir} \mid \mathbf{v}_{itr},\ t = 1,\ldots,\ T(i) = \prod_{t=1}^{T_i} P_{itr} \mid \mathbf{v}_{itr}.$$

The unconditional density would now be obtained by integrating the random terms out of the conditional distribution. We do this by simulation:

$$P_i = \frac{1}{R} \sum_{r=1}^{R} P_{ir} \mid (\mathbf{v}_{itr},\ t = 1,\ldots,T_i)$$

Note that in the random effects case, we are averaging over R replications in which the $T(i)$ observations are each a function of the same $\mathbf{v}_{ir}$. Thus, each replication in this case involves drawing a single random vector. In the AR1 case, each replication involves drawing a sequence of $T(i)$ vectors, $\mathbf{v}_{itr}$. Finally, the simulated log likelihood function to be maximized is

$$\log L = \sum_{i=1}^{N} \log P_i$$

$$= \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} P_{ir} \mid (\mathbf{v}_{itr},\, t = 1,...,T_i)$$

$$= \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{itr} \mid \mathbf{v}_{itr}$$

The derivatives must be approximated as well. The theoretical maximum is based on

$$\frac{\partial \log L_i}{\partial \boldsymbol{\Theta}} = \frac{1}{L_i} \frac{\partial L_i}{\partial \boldsymbol{\Theta}} = \mathbf{0}$$

where $\boldsymbol{\Theta}$ is the vector of all parameters in the model. Then,

$$\frac{\partial L_i}{\partial \boldsymbol{\Theta}} = \int_{\text{Range of } \mathbf{v}_{it}} g(\mathbf{v}_{it}, t = 1,...,T_i) \frac{\partial}{\partial \boldsymbol{\Theta}} \prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{it}' \mathbf{x}_{it}, \boldsymbol{\theta})\, d(\mathbf{v}_{it}, t = 1,...,T_i)$$

$$\frac{\partial}{\partial \boldsymbol{\Theta}} \prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{it}' \mathbf{x}_{it}, \boldsymbol{\theta}) = \sum_{t=1}^{T_i} \left[\frac{\partial P(y_{it}, \boldsymbol{\beta}_{it}' \mathbf{x}_{it}, \boldsymbol{\theta})}{\partial \boldsymbol{\Theta}} \right] \prod_{s \neq t} P(y_{is}, \boldsymbol{\beta}_{is}' \mathbf{x}_{is}, \boldsymbol{\theta})$$

$$= \prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{it}' \mathbf{x}_{it}, \boldsymbol{\theta}) \sum_{t=1}^{T_i} \left[\frac{\partial \log P(y_{it}, \boldsymbol{\beta}_{it}' \mathbf{x}_{it}, \boldsymbol{\theta})}{\partial \boldsymbol{\Theta}} \right]$$

Collecting terms once again, we obtain the approximation,

$$\frac{\partial \log L}{\partial \boldsymbol{\Theta}} = \sum_{i=1}^{N} \frac{1}{L_i} \frac{\partial L_i}{\partial \boldsymbol{\Theta}}$$

$$\approx \sum_{i=1}^{N} \frac{\left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{itr}' \mathbf{x}_{it}, \boldsymbol{\theta}) \right] \left[\sum_{t=1}^{T_i} \frac{\partial \log P(y_{it}, \boldsymbol{\beta}_{itr}' \mathbf{x}_{it}, \boldsymbol{\theta})}{\partial \boldsymbol{\Theta}} \right] \right\}}{\left\{ \frac{1}{R} \sum_{h=1}^{H} \left[\prod_{t=1}^{T_i} P(y_{it}, \boldsymbol{\beta}_{itr}' \mathbf{x}_{it}, \boldsymbol{\theta}) \right] \right\}}$$

Mechanics of computing the derivatives with respect to the low level parameters are given in the aforementioned technical document. They differ from model to model, so only the commonalities are shown here.

The Hessian is equally involved. We will only sketch the full derivation here. Return to the full gradient of the ith term in the log likelihood log likelihood - terms are summed over *i* to get the gradient and Hessian - the following is written in terms of the full parameter vector, including any ancillary parameters. The gradient is

$$\mathbf{g}_i = \frac{1}{P_i}\frac{1}{R}\sum_{r=1}^{R} P_{ir} \sum_{t=1}^{T_i} \frac{\partial \log P_{itr}}{\partial \boldsymbol{\Theta}}.$$

Let $\mathbf{H}_i$ denote the second derivatives matrix. Then,

$$\mathbf{H}_i = -\mathbf{g}_i \mathbf{g}_i' + \frac{1}{P_i}\frac{1}{R}\sum_{r=1}^{R} P_{ir}\left(\sum_{t=1}^{T_i} \frac{\partial \log P_{itr}}{\partial \boldsymbol{\Theta}}\right)\left(\sum_{t=1}^{T_i} \frac{\partial \log P_{itr}}{\partial \boldsymbol{\Theta}}\right)',$$
$$+ \frac{1}{P_i}\frac{1}{R}\sum_{r=1}^{R} P_{ir} \sum_{t=1}^{T_i} \frac{\partial^2 \log P_{itr}}{\partial \boldsymbol{\Theta}\, \partial \boldsymbol{\Theta}'}.$$

The only term which has not already appeared is the second derivatives matrix in the third part. Consider first the case of no autocorrelation and let $\boldsymbol{\mu}_k$ denote the vector of elements in $\boldsymbol{\Theta}$ that appear in $\beta_{k,itr}$. This derivative is obtained by differentiation of

$$\frac{\partial \log P_{itr}}{\partial \boldsymbol{\mu}_k} = \frac{\partial \log P_{itr}}{\partial \beta_{k,itr}} \frac{\partial \beta_{k,itr}}{\partial \boldsymbol{\mu}_k}.$$

which gives

$$\frac{\partial^2 \log P_{itr}}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_m'} = \frac{\partial \log P_{itr}}{\partial \beta_{k,itr}} \frac{\partial^2 \beta_{k,itr}}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_m'} + \frac{\partial \beta_{k,itr}}{\partial \boldsymbol{\mu}_k} \frac{\partial^2 \log P_{itr}}{\partial \beta_{k,itr} \partial \boldsymbol{\mu}_m'}.$$

In the absence of autocorrelation, the random parameters are linear in the underlying structural parameters, so the first of these two second derivatives is zero. Using this and our previous results, we obtain

$$\frac{\partial^2 \log P_{itr}}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_m'} = \left(\frac{\partial^2 \log P_{itr}}{\partial \beta_{k,itr} \partial \beta_{m,itr}}\right)\left(\frac{\partial \beta_{k,itr}}{\partial \boldsymbol{\mu}_k}\right)\left(\frac{\partial \beta_{m,itr}}{\partial \boldsymbol{\mu}_m'}\right).$$

The remaining complication in the preceding arises when there is autocorrelation, as in this case, the reduced form parameters are not linear in ρ_k. In this instance, the square of the first derivative is used as approximation to the second when the asymptotic covariance matrix is computed. (The algorithm used for estimation requires only first derivatives.)

Appendix B. Implementations of the Panel Data Estimators in Computer Software

The panel data estimators described in this article have all been implemented in *LIMDEP*, as listed in the table below. Other commercial packages also contain some of them. *Stata* contains a number of applications of the quadrature based procedures, the fixed effects count and logit models, and an extensive range of GEE formulations. *SAS* contains the logit and RE binomial models, some GEE models, and numerous variants of the linear model. Coelli's (1995) *FRONTIER* program contains all the panel estimators for the stochastic frontier model. *TSP* and *EViews* contain all variants of the linear model and a few of the quadrature based procedures for random effects. Gauss libraries in general circulation also provide some of the quadrature based random effects models and all variants of the linear regression model.

Model Class	**Fixed Effects**	**Random Effects**	**Random Parameters**	**Latent Class**
Linear Regression	•	• a	•	c
Binary Choice				
Probit	•	• a	•	•
Logit	•	• a	•	•
Complementary log log	•	• a	•	•
Gompertz	•	• a	•	•
Bivar. Probit/Selection		• a	•	
Multinomial Choice				
Multinomial Logit		• a	•	
Multinomial Probit		• b		
Ordered Probit/Logit	•	• a	•	•
Count Data				
Poisson Regression	•	• a	•	•
Negative Binomial	•	• a	•	•
Loglinear Models				
Exponential	•	• b	•	•
Gamma	•	• b	•	•
Weibull	•	• b	•	•
Inverse Gaussian	•	• b	•	•
Limited Dependent Variable				
Tobit	•	• a	•	•
Grouped data (censored)	•	• a	•	•
Truncated Regression	•	• b	•	•
Sample Selection	•	• b	•	
Survival and Frontier Models				
Weibull	•	• b	•	•
Exponential	•	• b	•	•
Loglogistic	•	• b	•	•
Lognormal	•	• b	•	•
Stochastic Frontier	•	• a	•	•

Notes: Any RP model produces an RE model by a random constant term. In the table, "a" denotes a model that can be estimated by standard REM techniques (GLS, quadrature) or by the simulation method with a random parameters formulation; "b" denotes a random effects model that can only be obtained by the simulated random parameters approach. The linear regression model can be fit with FEM by ML and LS, REM by GLS and simulated ML. The "c" indicates this model is not identified and therefore, not estimable. The binary choice models Complementary loglog and Gompertz can be fit with random effects by a random constant term in the RP model or by quadrature. The multinomial logit model is fit with random effects by random constant terms in the random parameters logit model in NLOGIT.