Estimating Econometric Models with Fixed Effects

William Greene*

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

April, 2001

Abstract

The application of nonlinear fixed effects models in econometrics has often been avoided for two reasons, one methodological, one practical. The methodological question centers on an incidental parameters problem that raises questions about the statistical properties of the estimator. The practical one relates to the difficulty of estimating nonlinear models with possibly thousands of coefficients. This note will demonstrate that the second is, in fact, a nonissue, and that in a very large number of models of interest to practitioners, estimation of the fixed effects model is quite feasible even in panels with huge numbers of groups. The models are fully parametric, and all parameters of interest are estimable.

Keywords: Panel data, fixed effects, computation.

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 from discussions with George Jakubson (who suggested one of the main results in this paper), Martin Spiess, 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 fixed effects model is a useful specification for accommodating individual heterogeneity in panel data. But, it has been problematic for two reasons. In most cases, the estimator is inconsistent owing to the incidental parameters problem. How serious this problem is in practical terms remains to be established - there is only a very small amount of received evidence - but the theoretical result is unambiguous. A second problem is purely practical. With current technology, the computation of the model parameters and appropriate standard errors, with all its nuisance parameters, appears to be impractical. This note focuses on the second of these, and shows that in a large number of interesting cases, the difficulty is only apparent. We will focus on a single result, computation of the estimator, and rely on some well known algebraic results to establish it. No formal statistical results are derived here or suggested with Monte Carlo results, as in general, the results are already known. The one statistical question noted above is left for further research.

The paper proceeds as follows. In Section 2, the general modeling framework is presented, departing from the linear model to more complicated specifications. The formal results for the estimator and computational procedures for obtaining appropriate standard errors are presented in Section 3. Section 4 suggests two possible new applications. Conclusions are drawn in Section 5.

2. Models with Fixed Effects

The linear regression model with fixed effects is

\[ y_{it} = \beta' x_{it} + \alpha_i + \delta_t + \epsilon_{it}, \quad t = 1, \ldots, T(i), \quad i = 1, \ldots, N, \]

\[ E[\epsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT(i)}] = 0, \]

\[ \text{Var}[\epsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT(i)}] = \sigma^2. \]

We have assumed the strictly exogenous regressors case in the conditional moments, [see Woolridge (1995)]. We have not assumed equal sized groups in the panel. The vector \( \beta \) is a set of parameters of primary interest, \( \alpha_i \) is the group specific heterogeneity. We have included time specific effects but, they are only tangential in what follows. Since the number of periods is usually fairly small, these can usually be accommodated simply by adding a set of time specific dummy variables to the model. Our interest here is in the case in which \( N \) is too large to do likewise for the group effects. For example in analyzing census based data sets, \( N \) might number in the tens of thousands. The analysis of two way models, both fixed and random effects, has been well worked out in the linear case. [See, e.g., Baltagi (1995) and Baltagi, et al. (2001).] A full extension to the nonlinear models considered in this paper remains for further research.

The parameters of the linear model with fixed individual effects can be estimated by the 'least squares dummy variable' (LSDV) or 'within groups' estimator, which we denote \( \hat{b}_{LSDV} \). This is computed by least squares regression of \( y_{it}^* = (y_{it} - \bar{y}_i) \) on the same transformation of \( x_i \) where the averages are group specific means. The individual specific dummy variable coefficients can be estimated using group specific averages of residuals. [See, e.g., Greene (2000, Chapter 14).] The slope parameters can also be estimated using simple first differences. Under the assumptions, \( \hat{b}_{LSDV} \) is a consistent estimator of \( \beta \). However, the individual effects, \( \alpha_i \), are each estimated with the \( T(i) \) group specific observations. Since \( T(i) \) might be small, and is, moreover, fixed, the estimator, \( \hat{a}_{LSDV} \), is inconsistent. But, the inconsistency of \( \hat{a}_{LSDV} \) is not transmitted to
b_{LSDV} because \( \bar{y}_i \) is a sufficient statistic. The LSDV estimator \( b_{LSDV} \) is not a function of \( a_{i,LSDV} \).

There are a few nonlinear models in which a like result appears. We will define a nonlinear model by the density for an observed random variable, \( y_{it} \):

\[
f(y_{it} | x_{i1}, x_{i2}, \ldots, x_{iT}) = g(y_{it}, \beta' x_{it} + \alpha_i, \theta)
\]

where \( \theta \) is a vector of ancillary parameters such as a scale parameter, an overdispersion parameter in the Poisson model or the threshold parameters in an ordered probit model. We have narrowed our focus to linear index function models. For the present, we also 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)]. However, it does appear that extension of the fixed effects model to dynamic models may well be practical. This, and multiple equation models, such as VAR's are 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 \( y_{it} \). 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.

The likelihood function for a sample of \( N \) observations is

\[
L = \prod_{i=1}^{N} \prod_{t=1}^{T(i)} g(y_{it}, \beta' x_{it} + \alpha_i, \theta).
\]

The likelihood equations,

\[
\frac{\partial \log L}{\partial \beta} = 0, \quad \frac{\partial \log L}{\partial \alpha_i} = 0, i = 1, \ldots, N, \quad \frac{\partial \log L}{\partial \theta} = 0,
\]

do not have explicit solutions for the parameter estimates in terms of the data and must, therefore, be solved iteratively. In principle, maximization can proceed simply by creating and including a complete set of dummy variables in the model. But, at some point, this approach becomes 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 \( 50,000 + K_{\beta} + K_{\theta} \) parameters. 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.

The proliferation of parameters is a practical shortcoming of the fixed effects model. The 'incidental parameters problem' is a methodological issue. If \( \beta \) and \( \theta \) were known, then, the solution for \( \alpha_i \) would be based on only the \( T(i) \) observations for group \( i \) (see below for an application). This implies that the asymptotic variance for \( \alpha_i \) is \( O(1/T(i)) \) and, since \( T(i) \) is fixed, \( \alpha_i \) is inconsistent. In fact, \( \beta \) is not known; in general in nonlinear settings, the estimator will be a function of the estimator of \( \alpha_i, a_{i,ML} \). Therefore \( b_{ML} \), MLE of \( \beta \) is a function of a random variable which does not converge to a constant as \( N \rightarrow \infty \), so neither does \( b_{ML} \). There is a small sample bias as well. The example is unrealistic, but Hsiao (1993, 1996) shows that in a binary logit model with a single regressor that is a dummy variable and a panel in which \( T(i) = 2 \) for all groups, the small sample bias is +100%. No general results exist for the small sample bias in more realistic settings. Heckman (1981) found in a Monte Carlo study of a probit model that the bias of the slope estimator in a fixed effects model was toward zero and 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. In many microeconometric applications, \( T(i) \) is considerably larger than this, so for practical purposes, there is good cause for optimism.

3
3. Computation of the Fixed Effects Estimator

In the linear case, regression using group mean deviations sweeps out the fixed effects. The slope estimator is not a function of the fixed effects which implies that it (unlike the estimator of the fixed effect) is consistent. There are a few analogous cases of nonlinear models that have been identified in the literature. Among them are the binomial logit model,

\[ g(y_{it}, \beta'x_{it} + \alpha_i) = \Lambda[(2y_{it} - 1)(\beta'x_{it} + \alpha_i)] \]

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

\[ g(y_{it}, \beta'x_{it} + \alpha_i) = (1/\lambda_{it})\exp(-y_{it}/\lambda_{it}), \lambda_{it} = \exp(\beta'x_{it} + \alpha_i), y_{it} \geq 0. \]

[See Munkin and Trivedi (2000) and Greene (2001).] In these models, there is a solution to the likelihood equation for \( \beta \) that is not a function of \( \alpha_i \). Consider the Poisson regression model with fixed effects - the result for the exponential model is essentially the same - for which

\[ \log L = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} -\exp(\alpha_i)\exp(\beta'x_{it}) + y_{it}(\beta'x_{it} + \alpha_i) - \log y_{it}! \]

The likelihood equation for \( \alpha_i, \partial \log L/\partial \alpha_i = 0, \) implies a solution

\[ \exp(\alpha_i) = \frac{\sum_{t=1}^{T(i)} y_{it}}{\sum_{t=1}^{T(i)} \exp(\beta'x_{it})}. \]

Thus, the maximum likelihood estimator of \( \beta \) is not a function of \( \alpha_i \). There are other models with loglinear conditional mean functions, however these are too few and specialized to serve as the benchmark case for a modeling framework. In the vast majority of cases of interest to practitioners, including those based on transformations of normally distributed variables such as the probit and tobit models, this method will be unusable.

Heckman and MacCurdy (1980) suggested a 'zig-zag' sort of approach to maximization of the log likelihood function, dummy variable coefficients and all. Consider the probit model. For known set of fixed effect coefficients, \( \alpha = (\alpha_1,...,\alpha_N)' \), estimation of \( \beta \) is straightforward. The log likelihood conditioned on these values (denoted \( a_i \)) would be

\[ \log L|a_1,...,a_N = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \Phi((2y_{it} - 1)(\beta'x_{it} + a_i)) \]
This can be treated as a cross section estimation problem since with known \( \alpha \), there is no connection between observations even within a group. With given estimate of \( \beta \) (denoted \( \mathbf{b} \)) the conditional log likelihood function for each \( \alpha_i \),

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

where \( z_{it} = \mathbf{b}'x_{it} \) is now a known function. Maximizing this function 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. There is no guarantee that this back and forth procedure will converge to the true maximum of the log likelihood function because the Hessian is not block diagonal. Whether either estimator is even consistent in the dimension of \( N \) (that is, of \( \beta \)) depends on the initial estimator being consistent, and it is unclear how one should obtain that consistent initial estimator. There is no maximum likelihood estimator for \( \alpha_i \) for any group in which the dependent variable is all 1s or all 0s, - the likelihood equation for \( \log L_i \) has no solution if there is no within group variation in \( y_{it} \). This feature of the model carries over to the tobit and binomial logit models, as the authors noted. In the Poisson and negative binomial models, any group which has \( y_{it} = 0 \) for all \( t \) contributes a 0 to the log likelihood function so its group specific effect is not identified either. Finally, irrespective of its probability limit, the estimated covariance matrix for the estimator of \( \beta \) will be too small, again because the Hessian is not block diagonal. The estimator at the \( \beta \) step does not obtain the correct submatrix of the information matrix.

Many of the models we have studied involve an ancillary parameter vector, \( \theta \). No generality is gained by treating \( \theta \) separately from \( \beta \), so at this point, we will simply group them in the single parameter vector \( \gamma = [\beta', \theta']' \). Denote the gradient of the log likelihood by

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

\[
\mathbf{g}_{\alpha_i} = \frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{T(i)} \frac{\partial \log g(y_{it}, \gamma, x_{it}, \alpha_i)}{\partial \alpha_i} \quad \text{(a scalar)}
\]

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

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

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

\[
\mathbf{H} = \begin{bmatrix}
\mathbf{H}_{\gamma\gamma} & \mathbf{h}_{y1} & \mathbf{h}_{y2} & \cdots & \mathbf{h}_{yN} \\
\mathbf{h}_{y1}' & h_{11} & 0 & \cdots & 0 \\
\mathbf{h}_{y2}' & 0 & h_{22} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{h}_{yN}' & 0 & 0 & 0 & h_{NN}
\end{bmatrix}
\]

where

\[
\mathbf{H}_{\gamma\gamma} = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \gamma, x_{it}, \alpha_i)}{\partial \gamma \partial \gamma'} \quad \text{(a } K_{\gamma} \times K_{\gamma} \text{ matrix)}
\]
\[ h_{ij} = \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \gamma, x_{it}, \alpha_i)}{\partial \gamma \partial \alpha_i} \] (N \times K_{\gamma} \times 1 \text{ vectors})

\[ h_{ii} = \sum_{t=1}^{T(i)} \frac{\partial^2 \log g(y_{it}, \gamma, x_{it}, \alpha_i)}{\partial \alpha_i^2} \] (N scalars).

Newton's method of maximizing the log likelihood produces the iteration

\[
\begin{pmatrix}
\gamma_k \\
\alpha_j_k
\end{pmatrix} = \begin{pmatrix}
\gamma_{k-1} \\
\alpha_j_{k-1}
\end{pmatrix} - H_{k-1}^{-1} g_{k-1} = \begin{pmatrix}
\gamma_k \\
\alpha_j_k
\end{pmatrix} + \begin{pmatrix}
\Delta_\gamma \\
\Delta_\alpha
\end{pmatrix}
\]

where subscript 'k' indicates the updated value and 'k-1' indicates a computation at the current value. Let \( H^{\gamma\gamma} \) denote the upper left \( K_\gamma \times K_\gamma \) submatrix of \( H^{-1} \) and define the \( N \times N \) matrix \( H^{\alpha\alpha} \) and \( K_\gamma \times N \) \( H^{\alpha\gamma} \) likewise. Isolating \( \gamma \), then, we have the iteration

\[
\gamma_k = \gamma_{k-1} - [H^{\gamma\gamma} g_\gamma + H^{\gamma\alpha} g_\alpha]_{k-1} = \gamma_{k-1} + \Delta_\gamma
\]

Using the partitioned inverse formula [e.g., Greene (2000, equation 2-74)], we have

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

The fact that \( H^{\alpha\alpha} \) is diagonal makes this computation simple. Collecting the terms,

\[
H^{\gamma\gamma} = \left[ H^{\gamma\gamma} - \sum_{i=1}^{N} \left( \frac{1}{h_{ii}} \right) h_{ij} h_{ji} \right]^{-1}
\]

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

\[
H^{\gamma\alpha} = -H^{\gamma\gamma} H^{\gamma\alpha} H^{\alpha\alpha}^{-1}
\]

As before, the diagonality of \( H^{\alpha\alpha} \) makes this straightforward. Combining terms, we find

\[
\Delta_\gamma = -H^{\gamma\gamma} (g_\gamma - H^{\gamma\alpha} H^{\alpha\alpha}^{-1} g_\alpha)
\]

\[
= - \left[ H^{\gamma\gamma} - \sum_{i=1}^{N} \left( \frac{1}{h_{ii}} \right) h_{ij} h_{ji} \right]^{-1} \left( g_\gamma - \sum_{i=1}^{N} \frac{g_{\alpha i}}{h_{ii}} h_{ji} \right)
\]
Turning now to the update for $\alpha$, we use the same results for the partitioned matrices. Thus,

$$\Delta_\alpha = - [H^{\alpha\alpha} g_\alpha + H^{\alpha\gamma} g_\gamma]_{i}.$$ 

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

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

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

Therefore,

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

$$= - H^{-1}_{\alpha\alpha} (g_\alpha + H_{\alpha\gamma} \Delta_\gamma).$$

Since $H_{\alpha\alpha}$ is diagonal,

$$\Delta_\alpha = \frac{1}{h_{ii}} (g_{\alpha i} + h_{ii} \Delta_\gamma).$$

Neither update vector requires storage or inversion of a $(K_{\gamma}+N) \times (K_{\gamma}+N)$ matrix; each is a function of sums of scalars and $K_{\gamma} \times 1$ vectors of first derivatives and mixed second derivatives.

The practical implication is that calculation of fixed effects models is a computation only of order $K_{\gamma}$. Storage requirements for $\alpha$ and $\Delta_\alpha$ are linear in $N$, not quadratic. Even for huge panels of tens of thousands of units, this is well within the capacity of even modest desktop computers. In experiments, we have found this method effective for probit models with 10,000 effects. (An analyst using this procedure for a tobit model reported success with nearly 15,000 coefficients.) (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 not an issue.)

The estimator of the asymptotic covariance matrix for the MLE of $\gamma$ is -$H^{\gamma\gamma}$, the upper left submatrix of -$H^{1}$. This is a sum of $K_{\gamma} \times K_{\gamma}$ matrices, and will be of the form of a moment matrix which is easily computed (see the application below). Thus, the asymptotic covariance matrix for the estimated coefficient vector is easily obtained in spite of the size of the problem. The asymptotic covariance matrix of $\alpha$ is

$$-(H_{\alpha\alpha} - H_{\alpha\gamma} H^{-1}_{\gamma\gamma} H_{\gamma\alpha})^{-1} = - H^{-1}_{\alpha\alpha} - H^{-1}_{\alpha\alpha} H_{\alpha\gamma} \{ H^{-1}_{\gamma\gamma} - H_{\gamma\alpha} H^{-1}_{\alpha\alpha} H_{\alpha\gamma} \}^{-1} H_{\gamma\alpha} H^{-1}_{\alpha\alpha}.$$ 

It is (presumably) not possible to store the asymptotic covariance matrix for the fixed effects estimators (unless there are relatively few of them). But, by expanding the summations where needed and exploiting the diagonality of $H_{\alpha\alpha}$, we find that the individual terms are

---

1 The iteration for the slope estimator is suggested in the context of the binary logit model by Chamberlain (1980, page 227). A formal derivation of $\Delta_\gamma$ and $\Delta_\alpha$ was presented by George Jakubson of Cornell University in an undated memo, "Fixed Effects (Maximum Likelihood) in Nonlinear Models."
\[ \text{Asy. Cov}[a_i, a_j] = -1(i = j) \frac{1}{h_{ii}} - \frac{1}{h_{jj}} \sum_{i=1}^{N} \frac{1}{h_{ii}} h_{ii} \left[ H^{-1} - \sum_{i=1}^{N} \frac{1}{h_{ii}} h_{ii} \right]^{-1} h_{jj} \]

Once again, the only matrix to be inverted is \( K \times K \), not \( N \times N \) (and, it is already in hand) so this can be computed by summation. It involves only \( K \times 1 \) 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;

\[ \text{Asy. Cov}[c, a'] = \text{Asy. Var}[c] \cdot H_{\alpha \alpha}^{-1}. \]

This involves \( N \times N \) and \( K \times N \) matrices, but it simplifies to

\[ \text{Asy. Cov}[c, a] = \text{Asy. Var}[c] \cdot \frac{h_{ii}}{h_{ii}} \]

4. Applications

To illustrate the preceding, we examine two applications, the binomial probit (and logit) model(s) and a sample selection model. (With trivial modification, the first of these will extend to many other models, as shown below.)

4.1. Binary Choice and Simple Index Function Models

For a binomial probit model with dependent variable \( z_{it} \),

\[ g(z_{it}, \beta'x_{it} + \alpha_i) = \Phi[(2z_{it} - 1)(\beta'x_{it} + \alpha_i)] = \Phi(q_{it} + r_{it}) = \Phi(a_{it}) \]

and

\[ \log L = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \Phi[q_{it}(\beta'x_{it} + \alpha_i)]. \]

Define the first and second derivatives of \( g(z_{it}, \beta'x_{it} + \alpha_i) \) with respect to \( (\beta'x_{it} + \alpha_i) \) as

\[ \lambda_{it} = q_{it} \frac{\Phi(a_{it})}{\Phi(a_{it})} \]

\[ \Delta_{it} = -a_{it} \lambda_{it} - \lambda_{it}^2, -1 < \Delta_{it} < 0. \]

The derivatives of the log likelihood for the probit model are

\[ \lambda_{it} = q_{it} \frac{\Phi(a_{it})}{\Phi(a_{it})} \]

\[ \Delta_{it} = -a_{it} \lambda_{it} - \lambda_{it}^2, -1 < \Delta_{it} < 0. \]
\[ g_{ai} = \sum_{t=1}^{T(i)} q_{it} \tilde{\lambda}_{it}, \]
\[ g_i = \sum_{j=1}^{N} \sum_{t=1}^{T(i)} q_{it} \tilde{\lambda}_{jt} x_{it}, \]
\[ h_{ii} = \sum_{t=1}^{T(i)} \Delta_{it}, \]
\[ h_{ji} = \sum_{t=1}^{T(i)} \Delta_{it} x_{it}, \]
\[ H_{ij} = \sum_{j=1}^{N} \sum_{t=1}^{T(i)} \Delta_{it} x_{it} x_{it}^\prime. \]

For convenience, let
\[ \Delta_i = \sum_{t=1}^{T(i)} \Delta_{it} \]
and
\[ \tilde{x}_i = h_{ii} / h_{ii} = \sum_{t=1}^{T(i)} \Delta_{it} x_{it} / \sum_{t=1}^{T(i)} \Delta_{it}. \]

Note that \( \tilde{x}_i \) is a weighted within group mean of the regressor vectors.

The update vectors and computation of the slope and group effect estimates follows the template given earlier. After a bit of manipulation, we find the asymptotic covariance matrix for the slope parameters is
\[
\text{Asy.Var}[b_{\text{MLE}}] = [-H_{ij}]^{-1} = -\left\{ \sum_{j=1}^{N} \left[ \sum_{t=1}^{T(i)} \Delta_{jt} \left( x_{jt} - \tilde{x}_j \right) \left( x_{jt} - \tilde{x}_j \right) \right] \right\}^{-1}
\]

The resemblance to the 'within' moment matrix from the analysis of variance context is notable and convenient. Inserting the parts and collecting terms produces

\[
\Delta_i = \left\{ \sum_{j=1}^{N} \left[ \sum_{t=1}^{T(i)} \Delta_{jt} \left( x_{jt} - \tilde{x}_j \right) \left( x_{jt} - \tilde{x}_j \right) \right] \right\}^{-1} \times \left\{ \sum_{j=1}^{N} \sum_{t=1}^{T(i)} q_{it} \tilde{\lambda}_{jt} \left( x_{jt} - \tilde{x}_j \right) \right\}
\]

and
\[
\Delta_{ij} = \left[ \sum_{t=1}^{T(i)} -q_{it} \tilde{\lambda}_{jt} / \Delta_i \right] + \tilde{x}_i \Delta_i.
\]

Denote the matrix in the preceding as
\[
V = [-H_{ij}]^{-1} = \text{Asy.Var}[b_{\text{MLE}}].
\]

Then,
\[
\text{Asy.Cov}[a_i, a_j] = -\frac{1}{\Delta_i} \mathbb{1}(i = j) + \tilde{x}_i V \tilde{x}_j = -\frac{1}{\Delta_i} \mathbb{1}(i = j) + s_{ij}
\]
Finally,
Each of these involves a moderate amount of computation, but can easily be obtained with existing software and, most important for our purposes, involves computations that are linear in $N$ and $K$. We note as well that the preceding extends directly to any other simple index function model, such as the binomial logit model [change derivatives $\lambda_{it}$ to $(1- \Lambda_{it})$ and $\Delta_{it}$ to $-\Lambda_{it}(1 - \Lambda_{it})$ where $\Lambda_{it}$ is the logit CDF] and the Poisson regression model [replace $\lambda_{it}$ with $(y_{it} - m_{it})$ and $\Delta_{it}$ with $-m_{it}$ where $m_{it} = \exp(\beta' x_{it} + \alpha_i)$]. Extension to models that involve ancillary parameters, such as the tobit model, are a bit more complicated, but not excessively so.

The preceding provides the estimator and asymptotic variances for all estimated parameters in the model. For inference purposes, note that the unconditional log likelihood function is computed. Thus, a test for homogeneity is straightforward using the likelihood ratio test. Finally, one would normally want to compute marginal effects for the estimated probit model. The conditional mean in the model is

$$E[z_{it} | x_{it}] = \Phi(\beta' x_{it} + \alpha_i)$$

so the slopes in the model are

$$\frac{\partial E[z_{it} | x_{it}]}{\partial x_{it}} = \beta \phi(\beta' x_{it} + \alpha_i) = \delta.$$

In many applications, marginal effects are computed at the means of the data. The heterogeneity in the fixed effects presents a complication. Using the sample mean of the fixed effects estimators, the estimator would be

$$Est.\frac{\partial E[z_{it} | x_{it}]}{\partial x_{it}} = b \phi(b' \bar{x} + \bar{a}) = d.$$

In order to compute the appropriate asymptotic standard errors for these estimates, we need the asymptotic covariance matrix for the estimated parameters. The asymptotic covariance matrix for the slope estimator is already in hand, so what remains is $Asy.Cov[b, \tilde{a}]$ and $Asy.Var[\tilde{a}]$. For the former,

$$Asy.Cov[b, \tilde{a}] = -\frac{1}{N} \sum_{i=1}^{N} V_i x_i = -V x$$

while, by summation, we obtain

$$Asy.Var[\tilde{a}] = \frac{1}{N^2} \left[ \sum_{i=1}^{N} \left( \frac{-1}{\Delta_i} \right) + \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} \right]$$

These would be assembled in a $(K+1)\times(K+1)$ matrix, say $V^*$. The asymptotic covariance matrix for the estimated marginal effects would be

$$Asy.Var[\delta] = GV^*G'$$
where the $K$ and one columns of $G$ are contained in

$$G = \phi(\beta^T \mathbf{x} + \alpha) \left[ I - (\beta^T \mathbf{x} + \alpha) \beta^T \mathbf{x} \right] - (\beta^T \mathbf{x} + \alpha) \beta$$

These results extend to any simple index function model including several discrete choice and limited dependent variable models. Likewise, the derivation for the marginal effects is actually generic, and extends to any model in which the conditional mean function is of the form $m(\beta^T \mathbf{x}_i + \alpha_i)$.

### 4.2. A Sample Selection Model

Several researchers have studied the application of Heckman’s (1979) sample selection model to panel data sets. [See Greene (2001) for a survey.] A formal treatment with fixed effects remains to be proposed. We formulate the model as follows:

1. **Probit selection equation:**
   
   $$z_{it} = 1(\beta^T \mathbf{x}_{it} + \alpha_i + u_{it}) > 0, \ E[u_{it}] = 0, \ Var[u_{it}] = 1.$$ 

2. **Regression equation - data observed when $z_{it} = 1$:**
   
   $$y_{it} = \gamma^T \mathbf{w}_{it} + \kappa_i + e_{it}, \ E[e_{it}] = 0, \ Var[e_{it}] = \sigma^2, \ Cov[e_{it}, u_{it}] = \rho \sigma.$$ 

The full log likelihood function for this model is

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \left\{ \Phi(-\beta^T \mathbf{x}_{it} - \alpha_i) \right\} + \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \left\{ \Phi(\beta^T \mathbf{x}_{it} + \alpha_i) \right\} \left( \frac{y_{it} - \gamma^T \mathbf{w}_{it} - \kappa_i}{\sigma} + \frac{\rho (\beta^T \mathbf{x}_{it} + \alpha_i)}{\sqrt{1-\rho^2}} \right) \frac{1}{\sigma}.$$ 

This involves all $K_\beta + K_\gamma + 2N + 2$ parameters. In principle, one could proceed in a fashion similar to that shown earlier, though with two full sets of fixed effects, the complication grows rapidly. We propose, instead, a two step approach which bears some resemblance to Heckman’s method for fitting this model by two step regression. At step 1, estimate the parameters of the probit model. At step 2, we estimate the remaining parameters, conditionally on these first step estimates. Each step uses the method of Section 3. This is the standard two step maximum likelihood estimator discussed in Murphy and Topel (1985) and Greene (2000, Chapter 4). The second step standard errors would then be corrected for the use of the stochastic parameter estimates carried in from the first step. We return to this consideration below.

The estimation of the parameters of the probit model, $\beta$ and $\alpha_i$, is discussed in the preceding section. For the second step, it is convenient to reparameterize the log likelihood function. Let $\delta = (1/\sigma) \gamma, \pi_i = \kappa_i/\sigma, \theta = 1/\sigma$, and $\tau = \rho / \sqrt{1-\rho^2}$ and let $d_{it} = \beta^T \mathbf{x}_{it} + \alpha_i$. With these simplifications, the log likelihood function reduces to

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \Phi(q_{it} d_{it}) + \sum_{i=1}^{N} \sum_{t=1}^{T(i)} \log \theta + \log \phi \left[ \sqrt{1 + \tau^2} (\theta y_{it} - \delta \cdot w_{it} - \pi_i) + \tau d_{it} \right]$$
For purposes of the conditional estimator, we treat $d_i$ as known. Thus, only the second half of the function need be considered. We now apply the procedure of Section 3. This provides the full set of estimates for the entire model. What remains is correction of the standard errors at the second step. The algebra at this step appears to be intractable, even with all the simplifications made thus far. We have had some success with bootstrapping (where $i = 1, \ldots, N$ is the sample). Further results await additional research.

5. Conclusions

The literature has treated the selection of the random and fixed effects models as a Hobson's choice. The computational difficulties and the inconsistency caused by the small $T(i)$ problem have made the fixed effects model unattractive. The practical issues may well be moot, but the methodological problem remains. Still, there is a compelling virtue of the fixed effects model as compared to the random effects model. The assumption of zero correlation between latent heterogeneity and included observed characteristics that is necessary in the random effects model is particularly restrictive.

With the exceptions noted earlier the fixed effects estimator has seen relatively little use in nonlinear models. The methodological issue noted above has been a major obstacle, though convincing results remain to be established. Hsiao's example is striking but quite unrealistic. Heckman's example based on $T(i) = 8$ gives cause for greater optimism, but more research on the question of how serious the small sample problems are would be useful. Modern data sets, particularly in finance, have quite large group sizes, often themselves larger than the $N$ in samples other researchers have used for fitting equally complex models. The practical difficulty of the fixed effects model 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 below. The computational aspects of fixed effects forms for many models are not complicated at all. We have implemented this computation in over twenty different modeling frameworks including: (1) linear regression; (2) binary choice (probit, logit, complementary log log, gompertz, bivariate probit with sample selection); (3) multinomial choice (ordered probit and ordered logit); (4) count data (poisson, negative binomial, zero inflated count models); (5) loglinear models (exponential, gamma, weibull, inverse Gaussian); (6) limited dependent variables (tobit, censored data, truncated regression, sample selection); (7) survival models (Weibull, exponential, loglogistic, lognormal); (8) stochastic frontier (half normal, truncated normal, heteroscedastic). In each instance, experiments involving thousands of coefficients were used to confirm the practicality of the estimators.

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 a result from Krailo and Pike (1984), it turns out that Chamberlain's binomial logit model is quite practical with $T(i)$ up to as high as 100. See, also, Maddala (1987).
References


