The Behavior of the Fixed Effects Estimator in Nonlinear Models

William Greene*

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

February, 2002

Abstract

The nonlinear fixed effects models in econometrics has often been avoided for two reasons one practical, one methodological. The practical obstacle relates to the difficulty of estimating nonlinear models with possibly thousands of coefficients. In fact, in a large number of models of interest to practitioners, estimation of the fixed effects model is feasible even in panels with very large numbers of groups. The more difficult, methodological question centers on the incidental parameters problem that raises questions about the statistical properties of the estimator. There is very little empirical evidence on the behavior of the fixed effects estimator. In this note, we use Monte Carlo methods to examine the small sample bias in the binary probit and logit models, the ordered probit model, the tobit model, the Poisson regression model for count data and the exponential regression model for a nonnegative random variable. We find three results of note: A widely accepted result that suggests that the probit estimator is actually relatively well behaved appears to be incorrect. Perhaps to some surprise, the tobit model, unlike the others, appears largely to be unaffected by the incidental parameters problem, save for a surprising result related to the disturbance variance estimator. Third, as apparently unexamined previously, the estimated asymptotic estimators for fixed effects estimators appear uniformly to be downward biased.

Keywords: Panel data, fixed effects, computation, Monte Carlo.

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 theoretical results in this paper), Paul Allison, Martin Spiess, Manuel Arellano 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

In the analysis of panel data with nonlinear models, researchers often choose among a number of unattractive alternatives. Notwithstanding the myriad unsolved problems of state persistence (dynamics) which are not considered here, the choice often reduces to one between a random effects and a fixed effects specification. The random effects model requires an unpalatable orthogonality assumption - consistency requires that the effects be uncorrelated with the included variables. The fixed effects model relaxes this assumption but is widely recognized to suffer from the incidental parameters problem (see Neyman and Scott (1948) and Lancaster (2000) - it is inconsistent because the asymptotic variance of the estimator of the main parameters is a function of the small and assumed fixed group size. Apparently, at least in some models that have been examined in detail, it is also biased in finite samples. How serious these problems are in practical terms remains to be established - there is only a very small amount of received evidence. A second problem is purely practical. With current technology, with some exceptions noted below, the computation of the model with all its nuisance parameters, with appropriate standard errors, appears to be impractical. But, in a large number of interesting cases, this difficulty is only apparent. Using some well known algebraic results, computation of the unconditional fixed effects estimator is quite feasible even in extremely large models.

Much of the received wisdom on the fixed effects estimator is derived from known results for the linear model which do not carry over to nonlinear models. (See Maddala (1987) and Baltagi (1995).) The infeasibility of using the results for the linear model, e.g., in the probit model, has produced some pessimism about the feasibility of the estimator. There now exists an extensive literature on semiparametric and GMM approaches for some of these panel data models with latent heterogeneity (e.g., Honore and Kyriazidou (2000). Among the practical limitations of these estimators are that although they may provide estimators of the primary slope parameters, they generally do not provide estimators for the full set of model parameters and thus

preclude computation of marginal effects, probabilities or predictions for the dependent variable. (Indeed, some estimation techniques which estimate only the slope parameters and only "up to scale" provide essentially only information about signs of coefficients and statistical significance of variables in the model.). In contrast, the fixed effects estimator is a full information estimator that, under its assumptions, provides results for all model parameters including the parameters of the heterogeneity. Thus, in spite of its several shortcomings, the fixed effects estimator has some virtues which suggest that it is worth a detailed look at its properties. This study will examine the behavior of the estimator in a large variety of nonlinear models.

While the results in the literature are unambiguous, they are qualitative in nature. The one piece of quantitative empirical evidence is Heckman's (1981) widely cited Monte Carlo study of the probit model in which he found that the small sample bias of the estimator appeared to be surprisingly small. However, his study examined a very narrow range of specifications, focused only on the probit model and, (as is crucial for this note), did not, in fact, examine a fixed effects model. Heckman analyzed the bias of the fixed effects *estimator* in a random effects *model* – his analysis included the orthogonality assumption noted earlier. In spite of its wide citation, Heckman's results are of limited usefulness for the case in which the researcher contemplates the fixed effects estimator precisely because the assumptions of the random effects model are inappropriate. Moreover, our results below are sharply at odds with Heckman's (even in his specification). We begin in Section 2 with a general specification for nonlinear models with fixed effects. Section 3 considers computation of the estimator (mechanical details for the practitioner are presented in the appendix). Section 4 contains a Monte Carlo study of the behavior of the estimator. We first consider the familiar question of asymptotic bias. We also examine the estimated standard errors produced by the estimator

We note at this point the main statistical conclusions of this paper. First, save for some documented cases, such as the Poisson model in which there actually is no incidental parameters problem, the skepticism about the estimator is appropriate. Save for the tobit model, we find that

the estimator is uniformly biased away from zero, and substantially so even when T is fairly large. Second, Heckman's encouraging results for the probit model appear to be incorrect. Third, the slope estimators in the tobit model do not appear to be affected by the incidental parameters problem. This is an unexpected result, but it must be tempered by a finding that the variance estimator is so affected. The variance estimator in the tobit model is a crucial parameter for inference and analysis purposes. On the other hand, the bias in the variance estimator appears to fall fairly quickly with increasing T. Finally, we find that in all cases in which the expected biases in the slope estimators emerges, it is away from zero, but at the same time, the estimated standard errors apear to be biased toward zero. Thus, in practical terms, the problem of incidental parameters is compounded. Some conclusions are drawn in Section 5.

2. Models with Fixed Effects

We consider a nonlinear model defined by the density for an observed random variable, y_{it},

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

where $\boldsymbol{\theta}$ is a vector of ancillary parameters such as a disturbance standard deviation, 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 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)]. 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}. The likelihood function for a sample of N observations is

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

The likelihood equations,

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad \frac{\partial \log L}{\partial \alpha_i} = 0, i = 1,..., N, \quad \frac{\partial \log L}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

generally 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. What makes this impractical is a nondiagonal second derivatives matrix (or some approximation to it) with possibly thousands of rows and columns. But, that consideration is misleading, a proposition we will return to presently.

The practical issues notwithstanding, there are some theoretical problems with the fixed effects model. The first is the proliferation of parameters already noted. The second is the 'incidental parameters problem.' If $\boldsymbol{\beta}$ 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 $O[1/T_i]$ and, since T_i is fixed, a_i is inconsistent. (Note, unlike other familiar cases, such as 'simultaneous equations bias,' the estimator is not inconsistent because it estimates some other parameter; it is inconsistent because its variance does not converge to zero as the sample size, N, increases. It is consistent in T_i.) The estimator of $\boldsymbol{\beta}$ will be a function of the estimator of α_i, $a_{i,ML}$. Therefore $\mathbf{b}_{ML}$, MLE of $\boldsymbol{\beta}$ is a function of a random variable which does not converge to a constant as $N \to \infty$, so neither does $\mathbf{b}_{ML}$. There may be a small sample bias as well. The example is unrealistic, but 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, Hsiao (1996) shows that the small sample bias is +100%. Abrevaya (1997) shows that Hsiao's result extends to more general binomial logit models as long as T_i continues to equal two; our Monte Carlo results below are consistent with this result. No general results exist for the small sample bias if T exceeds 2 or for other models. 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) 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. Indeed, for many microeconometric applications, T_i is considerably larger than this, so for practical purposes, there is good cause for optimism. We will reconsider these effects in the Monte Carlo investigation in Section 4.

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,

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

where $\Lambda(z) = \exp(z)/[1+\exp(z)]$. (See Chamberlain (1980) for the result and Greene (2000, Chapter 19) for discussion and practical details.) In this case, $\Sigma_t y_{it}$ is a minimal sufficient statistic for α_i, and estimation in terms of the conditional density provides a consistent estimator of $\boldsymbol{\beta}$. Three other commonly used models that have this property are the Poisson and negative binomial regressions for count data (see Hausman, Hall, and Griliches (1984)) and the exponential regression model for a continuous nonnegative variable,

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

(See Munkin and Trivedi (2000).) In all these cases, the conditional log likelihood,

$$\log L_c = \sum_{i=1}^{N} \log f\left(y_{i1}, y_{i2}, ..., y_{i,T_i} \,\middle|\, \left(\Sigma_{t=1}^{T_i} y_{it}\right), \mathbf{x}_{i1}, \mathbf{x}_{i2}, ...\right)$$

is a function of $\boldsymbol{\beta}$ that is free of the fixed effects. There are other similar models, such as the gamma regression model, however these are too few and specialized to serve as standard modeling platform. In the 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 MaCurdy (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, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)'$, estimation of $\boldsymbol{\beta}$ is straightforward. The log likelihood conditioned on these values (denoted a_i), would be

$$\log L|a_1,...,a_N = \sum_{i=1}^{N}\sum_{t=1}^{T_i} \log\Phi[(2y_{it}-1)(\boldsymbol{\beta}'\mathbf{x}_{it}+a_i)$$

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. With given estimate of $\boldsymbol{\beta}$ (denoted $\mathbf{b}$) the conditional log likelihood function for each α_i,

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

where $z_{it} = \mathbf{b}'\mathbf{x}_{it}$ 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. We note three problems with this approach: First, 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 $\boldsymbol{\beta}$) even if T is large, depends on the initial estimator being consistent, and it is unclear how one should obtain that consistent initial estimator. Second, in any group in which the dependent variable is all ones or all zeros, 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 feature of the model carries over to the tobit and binomial logit models, as the authors noted and to Chamberlain's conditional logit model and the Hausman et al. estimator of the Poisson model. In the Poisson and negative binomial models cases, any group which has $y_{it} = 0$ for all t contributes a zero to the log likelihood function so its group specific effect is not identified. Third, irrespective of its probability limit, the estimated standard errors for the estimator of $\boldsymbol{\beta}$ will be too small, again

because the Hessian is not block diagonal. The estimator at the β step does not obtain the correct submatrix of the information matrix.

Maximization of the log likelihood function can, in fact, be done by 'brute force,' even in the presence of possibly thousands of nuisance parameters. The strategy, which uses some well known results from matrix algebra is described in the appendix. Using these results, it is possible to compute directly both the maximizers of the log likelihood and the appropriate submatrix of the inverse of the analytic second derivatives for estimating asymptotic standard errors. The statistical behavior of the estimator is a separate issue, but it turns out that the practical complications are actually surmountable in many cases of interest to researchers.

4. Sampling Properties of the Fixed Effects Estimator

The literature contains few specific pieces of evidence on the behavior of this estimator. Andersen (1973) and Hsiao (1996) showed that in the binomial logit model with only the individual constants and a dummy variable on the right hand side with coefficient equal to 1.0, the unconditional maximum likelihood estimator of that coefficient will exhibit a persistent bias of +100% regardless of N. Heckman's analysis of the fixed effects probit estimator is summarized in Table 1. For the static case of interest here, his general results for the probit model with $N = 100$ and $T = 8$ suggest, in contrast to the evidence for the logit model, a slight *downward* bias in the slope estimator. The striking feature of his results is how small the bias seems to be even with T as small as 8.

We have been unable to replicate any of Heckman's results. Both his and our own results are shown in Table 1. Some of the difference can be explained by different random number generators. But, this would only explain a small part of the strikingly different outcomes of the experiments and not the direction. In contrast to Heckman, using his specification, we find that the probit estimator, like the logit estimator, is substantially biased away from zero when $T = 8$. Consistent with expectations, the bias is far less than the 100% that appears to appear when $T = 2$.

The table contains three sets of results. The first are Heckman's reported values. The second and third sets of results are our computations for the same study. Heckman based his conclusions on 25 replications. To control the possibility that some of the variation is due to small sample effects, we have redone the analysis using 100 replications. The results in the second and third row of each cell are strongly consistent with the familiar results for the logit model and with our additional results in the next section. The bias in the fixed effects estimator appears to be quite large, and, in contrast to Heckman's results, is away from zero in all cases. The proportional bias does not appear to be a function of the parameter value.

Table 1. Heckman's Monte Carlo Study of the Fixed Effects Probit Estimator
Experimental Design:
$Y_{it} = \sigma_\tau \tau_i + \beta z_{it} + \varepsilon_{it}$, $i = 1,\ldots,100$, $t = 1,\ldots,8$.
$\tau_i \sim N[0,1]$
$z_{it} = 0.1t + 0.5z_{i,t-1} + U_{it}$, $U_{it} \sim U[-0.5,0.5]$, $z_{i0} = 5 + 10.0U_{i0}$
$\varepsilon_{it} \sim N[0,1]$
$y_{it} = \mathbf{1}[Y_{it} > 0]$
(Note, initialization of z_{it} is given in Nerlove (1971).)

	β=1.0	**β=-0.1**	**β=-1.0**
$\sigma_\tau^2 = 3$	0.90[a] 1.286[b] 1.240[c]	-0.10 -0.1314 -0.1100	-0.94 -1.247 -1.224
$\sigma_\tau^2 = 1$	0.91 1.285 1.242	-0.09 -0.1157 -0.1127	-0.95 -1.198 -1.200
$\sigma_\tau^2 = 0.5$	0.93 1.213 1.225	-0.10 -0.1138 -0.1230	-0.96 -1.199 -1.185

[a]Reported in Heckman (1981), page 191.
[b]Mean of 25 replications
[c]Mean of 100 replications

For our purposes, there is an important shortcoming in the design of the foregoing experiment. The underlying model is not a fixed effects model; it is a random effects model. The signature feature of the fixed effects model is correlation between the effects and the included variables, and by construction, there is none between τ_i and z_{it} in the model above. As such, the foregoing does not give evidence on the point for which it is usually cited, that is, the small sample bias of the unconditional fixed effects estimator *of the fixed effects model*. More to the

point, if the researcher knows that the effects are not correlated with the included variables, then a random effects approach should be preferable, and the issue at hand becomes whether the normal distribution typically assumed is a valid assumption and what are the implications if it is not. Current technology provides a variety of useful approaches for random effects and random parameters models when it can be assumed that the effects and the included variables are orthogonal.

In this note, we will examine the behavior of the estimator in somewhat greater detail. We have as yet no evidence that Hsiao's result carries over to other models. In particular, we will examine two aspects of the fixed effects estimator:

- The small sample bias.
- Since the estimator is biased, it follows that the estimated asymptotic covariance matrix is as well. We will examine the extent to which the analytical estimator of the sampling variance agrees with the empirical, sampling variance of the estimator.

We will examine six index function models, the binomial logit, binomial probit, ordered probit, tobit, Poisson regression and exponential regression. These include binary, multinomial, censored/continuous, count, and nonnegative dependent variables. In the logit and Poisson cases, there exist sufficient statistics for the fixed effects parameters, so we can compare the sampling distributions of the two estimators. (In the Poisson case, the full fixed effects estimator is, in fact, consistent. We will return to this point below.) The experiment is designed as follows: All models are based on the same index function:

$$w_{it} = \alpha_i + \beta x_{it} + \delta d_{it}, \beta = \delta = 1,$$

where

$$x_{it} \sim N[0,1^2]$$

$$d_{it} = \mathbf{1}[x_{it} + h_{it} > 0] \text{ where } h_{it} \sim N[0,1^2]$$

$$\alpha_i = \sqrt{T}\bar{x}_i + a_i, a_i \sim N[0,1^2]$$

Thus, in all cases, we estimate the two coefficients on x_{it} and d_{it}, where both coefficients equal 1.0, and the fixed effects (which are not used or presented below). The correlations between the

variables are approximately 0.7 between x_{it} and d_{it}, 0.4 between α_i and x_{it} and 0.2 between α_i and d_{it}. The data generating processes examined are

Probit: $y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0]$,

Ordered Probit: $y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0] + \mathbf{1}[w_{it} + \varepsilon_{it} > 3]$,

Logit: $y_{it} = \mathbf{1}[w_{it} + v_{it} > 0]$, $v_{it} = \log[u_{it}/(1-u_{it})]$,

Tobit $y_{it} = \mathbf{1}[c_{it} > 0] \times c_{it}$, $c_{it} = w_{it} + \varepsilon_{it}$,

Poisson: $y_{it} = j \ni F(j|\lambda_{it}) < u_{it} < F(j+1|\lambda_{it})$, $F(z|\lambda_{it})$ = Poisson CDF, $\lambda_{it} = \exp(0.2w_{it})$,

Exponential: $y_{it} = \lambda_{it} \log u_{it}$, $\lambda_{it} = \exp(0.2w_{it})$,

where $\varepsilon_{it} \sim N[0,1^2]$ denotes a draw from the standard normal population and $u_{it} \sim U[0,1]$ denotes a draw from the standard uniform population. Models are fit with $T = (2, 3, 5, 8, 10, 20)$ and with $N = (100, 500, 1000)$. (Note that this includes Heckman's experiment.) Each model specification, group size, and number of groups is fit 200 times with random draws for ε_{it} or u_{it}. The conditioning data, x_{it}, d_{it} and α_i are held constant. The full set of parameters, including the dummy variable coefficients, are estimated using the results in the appendix. For each of the specifications listed, properties of the sampling distribution are estimated using the 200 observations on β and δ.

4.1. Small Sample Bias

Table 2 lists the means of the empirical sampling distribution for the six different estimators for the samples of 1,000 individuals. Thus, the influence of N should be minimized - these are large samples in this dimension. At this point, we are only interested in the mean of the sampling distribution as a function of T, so we use only the results based on the largest (N) samples. Note that in two cases, the conditional logit and conditional Poisson, the true bias is zero, as the estimator is not a function of the fixed effects. The bias of the fixed effects estimator in the binary and ordered choice models is large and persistent. Even at $T = 20$, we find fairly large biases. With $T = 2$, the Anderson/Hsiao result is clearly evident. Increasing the sample size

from 100 to 1,000 did little to remove this effect, but the increase in group size from 2 to 20 has a very large effect. We conclude that this is a persistent bias that can, indeed, be attributed to the "small *T* problem."

One quite striking effect in the tables is that it appears that the tobit fixed effects estimator is not biased at all. The result is all the more noteworthy in that in each data set, roughly 40 - 50% of the observations are censored. If none of the observations were censored, this would be a linear regression model, and the resultant OLS estimator would be the consistent LSDV estimator (by virtue of the Frisch-Waugh theorem). But, with roughly 40% of the observations censored, this is a quite unexpected result. However, the average of the 200 estimates of σ - the true value is also 1.0 - given in each cell for the tobit model shows that the incidental parameters problem shows up in a different place here. The estimated standard deviation is biased downward, though with a bias that does diminish substantially as *T* increases. This result is not inocuous. Consider estimating the marginal effects in the tobit model with these results. In general in the tobit model, for any variable in the model, $\delta_k = \partial E[y_i|\mathbf{x}_i]/\partial x_{ik} = \beta_k \times \Phi(\boldsymbol{\beta}'\mathbf{x}_i/\sigma)$ where $\Phi(z)$ is the cdf of the standard normal distribution. This is frequently computed at the sample means of the data. Based on our experimental design, the overall means of the variables would be zero for α_i and x_i and 0.5 for d_i. Therefore, the scale factor estimated, using the true values of the slope parameters as they are (apparently) estimated consistently, would be $\Phi(0.5/\hat{\sigma})$. The ratio of this value computed at the average estimate of σ to the value computed at $\sigma = 1$ is given in the third row of each cell in the table, where it can be seen that for small T, there is a substantive upward bias in the marginal effects. On the other hand, at T = 8 (Heckman's case), the tobit model appears to be essentially consistently estimated in spite of the incidental parameters issue.

The exponential and Poisson models displays no bias whatsoever. These models can be orthogonalized in the fashion derived by Lancaster (1997) so this is to be expected. This is a

useful practical result. In all these models, it is useful to have a full set of parameters for prediction and for analysis of marginal effects. The payoff to orthogonalizing the likelihood or basing estimation on the sufficient staitsitics in these cases is small in practical terms, and comes at the cost of losing the estimated fixed effects, themselves, needed for this second round of calculations. The brute force computation for these models is actually straightforward using the results in the appendix.

Table 2. Means of Empirical Sampling Distributions, N = 1000 Individuals Based on 200 Replications. Table entry is β,δ.

	T=2	*T*=3	*T*=5	*T*=8	*T*=10	*T*=20
	β δ	β δ	β δ	β δ	β δ	β δ
Logit	2.020, 2.027	1.698, 1.668	1.379, 1.323	1.217, 1.156	1.161, 1.135	1.069, 1.062
Logit-C[a]	0.994, 1.048	1.003, 0.999	0.996, 1.017	1.005, 0.988	1.002, 0.999	1.000, 1.004
Probit	2.083, 1.938	1.821, 1.777	1.589, 1.407	1.328, 1.243	1.247, 1.169	1.108, 1.068
Poisson	0.826, 0.761	0.978, 0.960	0.998, 0.995	0.991, 1.014	0.997, 1.006	1.003, 0.998
Poisson-C[b]	0.987, 1.018	0.995, 0.997	0.993, 1.015	1.002, 0.996	0.995, 1.015	1.000, 0.998
Tobit	0.981, 0.822	0.985, 0.991	0.997, 1.010	1.000, 1.008	1.001, 1.004	1.008, 1.001
σ	0.6444	0.7675	0.8642	0.9136	0.9282	0.9637
scale	1.13	1.07	1.04	1.02	1.01	1.02
Exponential	0.999, 0.962	0.998, 0.998	0.991, 0.993	0.998, 1.008	0.994, 1.012	0.997, 1.001
Ord. Probit	2.328, 2.605	1.592, 1.806	1.305, 1.415	1.166, 1.220	1.131, 1.158	1.058, 1.068

[a]Estimates obtained using the conditional likelihood function – fixed effects not estimated.
[b]Estimates obtained using Hausman et al's conditional estimator – fixed effects not estimated.

4.2. Estimates of the Asymptotic Standard Errors

In all the cases examined, a central issue is the extra variation induced in the parameter estimators by the presence of the inconsistent fixed effect estimators. Since the estimator, itself, is inconsistent, one should expect distortions in estimators of the asymptotic covariance matrix. Table 3 lists, for each model, the estimated asymptotic standard errors computed using the estimated second derivatives matrix and the empirical standard deviation based on the 200 replications in the simulation, using the N=100, T = 8 (Heckman's) group of estimators. The analytic estimator is obtained by averaging the 200 estimated asymptotic covariance matrices, then computing the square roots of the diagonal elements of the average matrix. The empirical estimator is the standard deviation of the 200 estimates obtained in the simulation. The latter should give a generally accurate assessment of the variation of the estimator while the former is,

itself, an estimator which is affected by the incidental parameters problem. There is clearly some downward bias in all the estimated standard errors. The implication is that as a general result, test statistics such as the Wald statistics (t ratios) will tend to be too large when based on the analytic estimator of the asymptotic variance – estimates are biased upward and apparently, standard errors are slightly biased downward. The two loglinear models seem to be unaffected by any of this; the empirical standard deviations and the analytic standard errors are essentially the same, again, as is to be expected..

Table 4. Estimated Standard Errors and Sample Standard Deviations of Sample Estimates

	Analytical		Empirical	
Model	β	δ	β	δ
Probit	0.2234	0.3008	0.2606	0.3254
Logit	0.2324	0.3697	0.2627	0.4312
Tobit	0.0692	0.1296	0.0800	0.1386
Ordered Probit	0.1281	0.2088	0.1487	0.2392
Poisson	0.2550	0.5290	0.2216	0.5228
Exponential	0.6765	1.2710	0.6483	1.3436

5. Conclusions

The computational difficulties and the inconsistency caused by the small T_i problem have made the fixed effects model unattractive and seem 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 considered.[1] Moreover, 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 issues may well be moot, but the methodological question of the incidental parameters problem remains. Still, there is a

[1]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

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.

The Monte Carlo results obtained here suggest a number of conclusions:

- As widely believed, the fixed effects estimator shows a large finite sample bias in discrete choice models when T is very small. The bias is persistent, but it does drop off rapidly as T increases to 3 and more. Heckman's widely cited result for the probit model appears to be incorrect, however. The discrepancy does not appear to be a function of the mechanism used to generate the exogenous variables. Heckman used Nerlove's (1971) dynamic model whereas we used essentially a random cross section. Results were similar for the two cases.
- The estimator shows essentially no bias in the slope estimators of the tobit model. But, the small sample bias appears to show up in the estimate of the disturbance variance.
- All the estimators save for the Poisson and exponential appear to underestimate the correct asymptotic variance. Thus, inference based on the conventional standard errors could be problematic.

Finally, at several points in the preceding, it was noted that one purpose for pursuing this estimator is that for better or worse, it does provide estimates of the fixed effects parameters. How good these estimators might be is an unanswered question, since, in the end, each is a function of only T observations. Superficially, one might argue that some information is better than none. On the other hand, for purposes of analyzing marginal effects, as suggested in the appendix, the average of these estimators might be useful and, depending on the assumptions underlying the generation of the effects, might well be a consistent (in $1/N$) estimator of a useful quantity. This does not imply, however, that one should simply ignore the issue and fit the

---

with T_i up to as high as 100. See, also, Maddala (1987). The Monte Carlo study done here involves fitting Chamberlain's model with $T = 20$.

original model with one constant. The fixed effects linear regression model can provide sufficient guidance on how that would affect the resulting estimates.

References

Abrevaya, J. 1997. The equivalence of two estimators of the fixed-effects logit model. *Economics Letters* **55, 1:** 41-44.

Ahn, S. and P. Schmidt. 1995. Efficient estimation of models for dynamic panel data. *Journal of Econometrics* **68**: 3-38.

Andersen, E. 1973. *Conditional Inference and Models for Measuring.* Mentalhygiejnisk Forsknings Institut: Copenhagen.

Arellano, M. and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**: 277-297.

Arellano, M. and O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* **68:** 29-51.

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

Baltagi, B., Song, S. and Jung, B. 2001. The unbalanced nested error component regression model. *Journal of Econometrics* **101**: 357-381.

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* **47**: 225-238.

Greene, W. 1993. *Econometric Analysis*, 2^{nd} *ed*. Prentice Hall: Englewood Cliffs.

Greene, W. 2000. *Econometric Analysis*, 4^{th} *ed*. Prentice Hall: Englewood Cliffs.

Hausman, J., B. Hall and Z. Griliches. 1984. Econometric models for count data with an application to the patents - R&D relationship. *Econometrica* **52**: 909-938.

Heckman, J. 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications*, Manski, C. and McFadden D. (eds). MIT Press: Cambridge.

Heckman, J. and MaCurdy, T. 1981. A life cycle model of female labor supply. *Review of Economic Studies* **47**: 247-283.

Holtz-Eakin, D. 1988. Testing for individual effects in autoregressive models. *Journal of Econometrics* **39**: 297-307.

Holtz-Eakin, D., W. Newey and S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica* **56**: 1371-1395.

Holtz-Eakin, D., W. Newey and S. Rosen. 1989. The revenues-expenditures nexus: evidence from local government Data. *International Economic Review* **30**: 415-429.

Honore, B. and T. Kyriazidou. 2000. Panel data discrete choice models with lagged dependent variables. *Econometrica* **68, 4**: 839-874.

Hsiao, C. 1996. Logit and probit models. In *The Econometrics of Panel Data: Handbook of Theory and Applications, Second Revised Edition,* Matyas, L. and Sevestre, P. (eds.). Kluwer Academic Publishers: Dordrecht.

Krailo, M. and M. Pike. 1984. Conditional multivariate logistic analysis of stratified case-control studies. *Applied Statistics* **44**, 95-103.

Lancaster, T. 2000. The incidental parameters problem since 1948. *Journal of Econometrics*, **95**: 391-414.

Maddala, G. 1987. Limited dependent variable models using panel data. *Journal of Human Resources* **22**: 307-338.

Nerlove, M. 1971. Further evidence on the estimation of dynamic economic relations from a time series of cross sections., *Econometrica* **39**: 359-382.

Neyman, J. and E. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* **16**: 1-32.

Munkin, M. and P. Trivedi. 2000. Econometric analysis of a self selection model with multiple outcomes using simulation-based estimation: An application to the demand for health care. Manuscript, Department of Economics, Indiana University.

Orme, C. 1999. Two-step inference in dynamic non-Linear panel data models," Manuscript, School of Economic Studies, University of Manchester.

Prentice, R. and L. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**: 57-67.

Woolridge, J. 1995. Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* **68**: 115-132.

Appendix: Computation of the Fixed Effects Estimator in Nonlinear Models

Many of the models we have studied involve an ancillary parameter vector, $\boldsymbol{\theta}$. 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 $\boldsymbol{\gamma} = [\boldsymbol{\beta}', \boldsymbol{\theta}']'$.

Denote the gradient of the log likelihood by

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

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

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

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

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

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

$$\mathbf{h}_{\gamma I} = \sum_{t=1}^{T_i} \frac{\partial^2 \log g(y_{it}, \boldsymbol{\gamma}, \mathbf{x}_{it}, \alpha_i)}{\partial \boldsymbol{\gamma} \partial \alpha_i} \text{ (} N K_\gamma \times 1 \text{ vectors)}$$

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

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

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

where subscript 'k' indicates the updated value and 'k-1' indicates a computation at the current value. 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

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

Using the partitioned inverse formula [e.g., Greene (2003, equation A-74)], we have

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

Since $\mathbf{H}_{\alpha\alpha}$ is diagonal,

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

Thus, the upper left part of the inverse of the Hessian can be computed by summation of vectors and matrices of order K_γ. Once again using the partitioned inverse formula,

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

Combining terms, we find that

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

Turning now to the update for $\boldsymbol{\alpha}$, we use the same results for the partitioned matrices. Thus,

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

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

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

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

Therefore,

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

Since $\mathbf{H}_{\alpha\alpha}$ is diagonal,

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

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.[2] The practical implication is that calculation of fixed effects models is a computation only of order K_γ. Storage requirements for α and Δ_α 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 of the current vintage.

The estimator of the asymptotic covariance matrix for the MLE of γ is $-\mathbf{H}^{\gamma\gamma}$, the upper left submatrix of $-\mathbf{H}^{-1}$. Since this is a sum of $K_\gamma\times K_\gamma$ matrices, 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 $\mathbf{a}$ is

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

The individual terms are

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

while

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

so

[2] The iteration for the slope estimator is suggested in the context of a binary choice model in Chamberlain (1980, page 227). A formal derivation of Δ_γ and Δ_α was given to the author by George Jakubson of Cornell University in an undated memo, "Fixed Effects (Maximum Likelihood) in Nonlinear Models." A similar result appears in Prentice and Gloeckler (1978).

$$Asy.Cov[\mathbf{c},a_i] = Asy.Var[\mathbf{c}]\times\left(\frac{\mathbf{h}_{\gamma i}}{h_{ii}}\right)$$

To illustrate the preceding, consider the binomial probit (and logit) model(s). With trivial modification, the results will extend to many other models, as shown below.)[3] For a binomial probit model with dependent variable z_{it},

$$g(z_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) = \Phi[(2z_{it} - 1)(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)] = \Phi(q_{it}\, r_{it}) = \Phi(a_{it})$$

and

$$\log L = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\log\Phi[q_{it}(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)].$$

Define the following first and second derivatives of $\log g(z_{it}, \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$:

$$\lambda_{it} = q_{it}\frac{\phi(a_{it})}{\Phi(a_{it})},\ \Delta_{it} = -a_{it}\lambda_{it} - \lambda_{it}^2,\ -1 < \Delta_{it} < 0.$$

The derivatives of the log likelihood for the probit model are

$$g_{\alpha i} = \sum_{t=1}^{T_i} q_{it}\lambda_{it},\quad \mathbf{g}_{\gamma} = \sum_{i=1}^{N}\sum_{t=1}^{T_i} q_{it}\lambda_{it}\mathbf{x}_{it},$$

$$h_{ii} = \sum_{t=1}^{T_i}\Delta_{it},\qquad \mathbf{h}_{\gamma i} = \sum_{t=1}^{T_i}\Delta_{it}\mathbf{x}_{it},\ \mathbf{H}_{\gamma\gamma} = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\Delta_{it}\mathbf{x}_{it}\mathbf{x}_{it}'.$$

For convenience, let $\Delta_i = h_{ii}$ and

$$\bar{\mathbf{x}}_i = \mathbf{h}_{\gamma i} / h_{ii} = \sum_{t=1}^{T_i}\Delta_{it}\mathbf{x}_{it} \Big/ \sum_{t=1}^{T_i}\Delta_{it}$$

Note that $\bar{\mathbf{x}}_i$ 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

$$\text{Asy.Var}[\mathbf{b}_{MLE}] = [-\mathbf{H}_{\gamma\gamma}]^{-1} = -\left\{\sum_{i=1}^{N}\left[\sum_{t=1}^{T(i)}\Delta_{it}\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)\right]\right\}^{-1}$$

[3] We assume in the following that none of the groups have y_{it} always equal to 1 or 0. In practice, one would have to determine this as part of the estimation effort. It should be noted for the practitioner that this condition is not trivially obvious during estimation. The usual criteria for convergence, such as small Δ_α will appear to be met even in the presence of degenerate groups while the associated α_i is still finite.

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

$$\Delta_\gamma = \left\{\sum_{i=1}^{N}\left[\sum_{t=1}^{T_i}\Delta_{it}\left(\mathbf{x}_{it}-\overline{\mathbf{x}}_i\right)\left(\mathbf{x}_{it}-\overline{\mathbf{x}}_i\right)\right]\right\}^{-1} \times \left\{\sum_{i=1}^{N}\left[\sum_{t=1}^{T(i)} q_{it}\lambda_{it}\left(\mathbf{x}_{it}-\overline{\mathbf{x}}_i\right)\right]\right\}$$

and

$$\Delta_{\alpha i} = \left[\sum_{t=1}^{T_i}\left(-q_{it}\lambda_{it}/\Delta_i\right)\right] + \tilde{\mathbf{x}}_i'\Delta_\gamma$$

Denote the matrix in the preceding as

$$\mathbf{V} = -[\mathbf{H}_{\gamma\gamma}]^{-1} = \text{Asy.Var}[\mathbf{b}_{MLE}].$$

Then,

$$\text{Asy.Cov}[a_i,a_j] = \frac{-\mathbf{1}(i=j)}{\Delta_i} + \overline{\mathbf{x}}_i'\mathbf{V}\overline{\mathbf{x}}_j = \frac{-\mathbf{1}(i=j)}{\Delta_i} + s_{ij}$$

and

$$\text{Asy.Cov}[\mathbf{b}_{MLE},a_i] = -\mathbf{V}\overline{\mathbf{x}}_i.$$

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 λ_{it} to $(1- \Lambda_{it})$ and Δ_{it} to $-\Lambda_{it}(1 - \Lambda_{it})$ where Λ_{it} is the logit CDF] and the Poisson regression model [replace λ_{it} with $(y_{it} - m_{it})$ and Δ_{it} with $-m_{it}$ where $m_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$]. 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

$$E[z_{it} \mid \mathbf{x}_{it}] = \Phi(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$$

so the slopes in the model are

$$\frac{\partial E[z_{it} \mid \mathbf{x}_{it}]}{\partial \mathbf{x}_{it}} = \boldsymbol{\beta}\, \phi(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i) \;=\; \boldsymbol{\delta}.$$

In many applications, marginal effects are computed at the means of the data. The heterogeneity in the fixed effects presents a new complication. One might compute the function at the means of the data and the sample mean of the fixed effects estimators. Thus, the estimator would be

$$Est.\frac{\partial E[z_{it} \mid \mathbf{x}_{it}]}{\partial \mathbf{x}_{it}} = \mathbf{b}\, \phi(\mathbf{b}'\bar{\mathbf{x}} + \bar{a}) \;=\; \mathbf{d}$$

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[$\mathbf{b}, \bar{a}$] and Asy.Var[$\bar{a}$]. For the former,

$$\text{AsyCov}[\mathbf{b}, \bar{a}\,] \;=\; \frac{-1}{N}\sum_{i=1}^{N} \mathbf{V}\bar{\mathbf{x}}_i = -\mathbf{V}\bar{\bar{\mathbf{x}}}$$

while, by simple summation, we obtain

$$\text{Asy.Var}[\,\bar{a}\,] \;=\; \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{-1}{\Delta_i} \;+\; \sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}\right]$$

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

$$\text{Asy.Var}[\boldsymbol{\delta}] \;=\; \mathbf{GV^*G'}$$

where the K and one columns of **G** are contained in

$$\mathbf{G} \;=\; \phi(\boldsymbol{\beta}'\bar{\mathbf{x}} + \bar{\alpha})\left[\mathbf{I} - (\boldsymbol{\beta}'\bar{\mathbf{x}} + \bar{\alpha})\boldsymbol{\beta}\bar{\mathbf{x}}' \;\middle|\; -(\boldsymbol{\beta}'\bar{\mathbf{x}} + \bar{\alpha})\boldsymbol{\beta}\right]$$

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 $m(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$.