ELSEVIER Journal of Econometrics 64 (1994) 165-182

JOURNAL OF Econometrics

A two-stage estimator for probit models with structural group effects

George J. Borjas, Glenn T. Sueyoshi*

NBER and University of California, San Diego, La Jolla, CA 92093-0508, USA

(Received October 1991; final version received August 1993)

Abstract

This paper outlines a two-stage technique for estimation and inference in probit models with structural group effects. The structural group specification belongs to a broader class of random components models. In particular, individuals in a given group share a common component in the specification of the conditional mean of a latent variable. For a number of computational reasons, existing random effects models are impractical for estimation and inference in this type of problem. Our two-stage estimator provides an easily estimable alternative to the random effects specification. In addition, we conduct a Monte Carlo simulation comparing the performance of alternative estimators, and find that the two-stage estimator is superior – both in terms of estimation and inference – to traditional estimators.

Key words: Random effects; Fixed effects; Panel data; Probit; Nonlinear regression
JEL classification: C23; C33; C63

1. Introduction

This paper outlines a two-stage technique for estimation and inference in probit models with structural group effects. The group effects, linear regression model has become increasingly popular in applied research. In this class of models, individuals belonging to a given group – for example, a given ethnic

* Corresponding author.

We have benefited greatly from discussions with Stephen Portnoy and Xiaohong Chen. All errors are our own.

SSDI 0304407693015827

classification or regional location – share a common component in the specification of a conditional mean. Often, researchers are interested in the determinants of the group components, and relate them to variables shared by the individuals in a group.[1]

The structural group effects specification may be viewed as a member of a broader class of random components models (Moulton, 1986). Consequently, the theory of estimation and inference for the linear regression, group effects model is well established; in particular, Hsiao (1986) and Judge et al. (1988), among others, provide surveys of the relevant literature outlining efficient GLS estimation techniques for linear random effects models.

There do not exist, to our knowledge, comparable analyses for latent variables models which are readily applicable to structural group effects specifications. In this paper, we extend the existing analysis of the linear regression model to the latent variable probit specification.[2] We develop and analyze a computationally tractable, two-stage estimator for individual and group level parameters which parallels Amemiya's (1978) two-stage estimator for the linear specification. In the first stage, we estimate a probit model which pools the observations across groups, but accounts for common effects by including dummy variables to allow for group-specific intercepts. Under reasonable regularity conditions, we avoid the Neyman and Scott (1948) incidental parameters problem and are able to estimate consistently the coefficients of dummy variables which represent the group effects. In the second stage, we fit the estimated group-specific intercepts to group level variables, employing GLS techniques to correct for nonspherical errors. We analyze the asymptotic behavior of the second-stage estimator for both the case where the number of groups is fixed and the case where the number of groups tends to infinity.

In Section 2, we outline the latent variable group effects regression specification and discuss existing estimation techniques for these models. Section 3 provides a small simulation comparing the performance of alternative estimators under both normal and alternative distributional assumptions for the group error and for various sample sizes. Section 4 provides some concluding comments.

---

[1] Recent empirical examples of this framework are provided by such diverse studies as Blanchflower and Oswald (1990), which analyzes the impact of SMSA-specific unemployment rates on the earnings of workers; Borjas (1992), which analyzes the factors underlying earnings differentials across ethnic groups; Case and Katz (1991), which analyzes the impact of neighborhood effects on socioeconomic outcomes; and Rauch (1992), which analyzes the impact of the average SMSA level of the human capital stock on the earnings of workers.

[2] While we focus our analysis on the probit model, the techniques for applying our results to other nonlinear models are obvious.

2. Probit models with group effects

2.1. Model specification

Consider a latent variables regression specification with random effects resulting from a group-specific error term,

$$Y_{ij}^* = X_{ij}'\beta + Z_j'\gamma + \varepsilon_{ij} + u_j \tag{1}$$

for groups $j = 1, \ldots, J$ and individuals $i = 1, \ldots, N_j$, for a total of $\sum_j N_j = N$ observations. X_{ij} is a K vector of explanatory variables, Z_j is an M-dimensional vector of explanatory variables common to members of group j, β and γ are conformable vectors of parameters. The ε are i.i.d. normal errors which are independent of the i.i.d. group errors u; both the ε and the u are assumed to be orthogonal to the Z and X. We begin with the assumption that u_j is distributed normally; this assumption will be relaxed subsequently.[3] In the case where Y_{ij}^* is fully observable, this specification is analogous to Amemiya's (1978) specification of the random effects model. The primary assumption that we make about the sample design is that the number of individuals per group (N_j) approaches infinity along with the total sample N.

Economic theory often generates models which describe how Y_{ij}^* should vary among groups as a function of variables Z characterizing the background or the opportunity set facing each group. Given our specification above, we may define a 'group effect' dummy variable, $d_j = Z_j'\gamma + u_j$, which represents the total effect of membership in group j on the latent Y_{ij}^*. Note that our specification of the group effect as a linear function of group-specific variables Z and an error term u provides a structural interpretation of the dummy variable d in terms of group-specific characteristics.

We do not observe Y_{ij}^* directly, but instead observe the indicator variable $Y_{ij} = 1\ (Y_{ij}^* > 0)$. The observable data consist of sets of observations on (Y_{ij}, X_{ij}', Z_j') for individuals i and groups j. In a slightly different context, data of this form have been termed a random effects probit model by Heckman and Willis (1975). Given the usual normality assumptions for u and ε, the errors $w = u + \varepsilon$ are multivariate normal with mean 0 and block-diagonal covariance matrix Σ.

There are three primary techniques for estimation of random effects probit models: pooled probit (Maddala, 1986, p. 317; Robinson, 1982), random effects (Heckman and Willis, 1975; Butler and Moffitt, 1982), and minimum distance

[3] In this formulation of the group effects model, correlation between the group effects and the explanatory variables X results from correlation between the X and the Z. See Mundlak (1978), Chamberlain (1984), and Hausman and Taylor (1981) for alternative specifications.

(Chamberlain, 1984).[4] Each of these approaches is unsuitable for the groups effects specification. The pooled probit will generate consistent and asymptotically normal, but inefficient, estimates of the parameters of interest, with standard statistical packages providing incorrect estimates of the standard errors. Computing the correct variance matrix for the parameter estimates for the pooled specification appears to us to be very difficult.[5] The Heckman and Willis random effects specification is theoretically appropriate for large J, but the performance of the estimator may not be satisfactory for large N_j since the model requires numeric integration over a term involving the product of cumulative normals for all group members. Large group sample sizes are likely to have numerical integrals which are at best, quite unstable, and at worst, involve integration over values which are smaller than machine precision.[6] The Chamberlain two-stage approach simply appears to be impractical for group effects data for a number of reasons. The first-stage, consistent estimation procedure requires the estimation of separate probit models for common observations across groups. In addition to the computational burden and obvious difficulties in justifying standard asymptotic results, there are conceptual difficulties since there is no natural ordering of observations for group effects data.[7]

2.2. *A two-step estimator*

Our approach takes advantage of the fact that as the number of individuals in each group increases, we are able to avoid the dimensionality problem associated with estimating models conditional on group membership. We propose estimating β from a fixed effects probit specification and γ via a second-stage GLS regression of the consistently estimated fixed effects on the group level variables. This approach is a natural extension of the Amemiya (1978) two-step estimator for the linear random effects model. For expository reasons, we consider a balanced sample with J groups and $N/J \to \infty$ observations per

[4] Alternative assumptions about the error terms may generate additional approaches. For example, if the ε are distributed as a logistic, conditional likelihood estimation (Anderson, 1970; Chamberlain, 1980) is possible.

[5] See, for example, the technical appendix in Robinson (1982) which includes formulae for computing the variance matrix in the case where the residuals in the model follow an AR(1) process.

[6] For a hypothetical sample of 500 observations per group assuming a generous likelihood contribution of 0.5 for each observation in a group, the value of the integrand is $e^{500 \times \log(0.5)} \approx e^{-346.6}$, which is well below the capabilities of most existing computer precision. For as few as 30 observations in a group, the value is on the order of $e^{-27.7}$. As a rough approximation, it would appear that group sizes over 50 may create significant instabilities if the model has low predictive power.

[7] We consider a modified Chamberlain approach which estimates models stratified by group in Section 2.2.3. Avery, Hansen, and Hotz (1983) propose a more general, but related generalized method of moments estimator.

group. We begin with the straightforward case of J fixed, and defer consideration of the case $J \to \infty$ until Section 2.2.3.

2.2.1. First-stage fixed effects

Suppose that we estimate a standard probit model, conditioning on group membership by including dummy variables for each of the J groups. We are interested in estimates of the true $(K + J)$-dimensional parameter vector $\theta'_0 = (\beta'_0, d'_0)$. Note that since $N/J \to \infty$ and J fixed, all of the components of the parameter vector θ are structural in the sense of Neyman and Scott.

Under appropriate regularity conditions, $\hat{\theta}$, the maximum likelihood estimator is a consistent root for θ_0 the true parameter vector and $\sqrt{N}$ asymptotically normal distributed with asymptotic variance given by $V = A^{-1}$, where $A = \lim_{N\to\infty} N^{-1}\mathrm{E}(-\nabla^2 l(\theta_0))$ and $\nabla^2 l(\theta)$ is the $((K + J) \times (K + J))$ matrix of log-likelihood second derivatives. For the probit model, Amemiya (1985, pp. 270–273) provides sufficient conditions for standard asymptotic results to obtain. Abstracting from the usual considerations regarding collinearity of the X, the force of the restriction on the information matrix necessary for asymptotic results requires $N_j/N \to \pi_j > 0$.[8] Put differently, so long as there are variations in the binary responses within a group and the group sizes all tend to infinity, we can avoid the incidental parameters problem and apply ML to the fixed effects model.[9]

2.2.2. Second-stage GLS estimation

Using the estimated $\hat{d}$, we specify a second-stage regression model as $\hat{d} = Z\gamma + w$; $w = u + (\hat{d} - d_0)$ where $(\hat{d} - d_0)$ is $o_p(1)$ and $\sqrt{N/J}(\hat{d} - d_0)$ is $O_p(1)$ so that $w \to u$ as $N \to \infty$. Then asymptotically, the second-stage model which regresses $\hat{d}$ on Z is a simple linear regression specification with normally distributed spherical errors and variance matrix $\sigma_u^2 I_J$ (applying results from Randles, 1982, and Pierce, 1982). Thus, a second-stage OLS estimator for γ employing $\hat{d}$ will asymptotically be both unbiased and normally distributed,

[8] This condition is satisfied trivially for the balanced design since $N_j = N/J$. As is the case for Berkson's minimum chi-square methods, if this condition fails for some j, we can respecify by ignoring the observations corresponding to that group (see, for further discussion, Amemiya, 1985, p. 276).

[9] See Heckman (1981a) for further discussion and Monte Carlo evidence that the estimator may be well-behaved for N/J as small as eight. Computationally, large J may pose some difficulties since there is no approach for removing probit fixed effects via 'differencing' or conditional likelihood (Anderson, 1970; Chamberlain, 1984). For large problems, we recommend a simplification of the usual ML estimation procedures which employs the recursive updating techniques described by Hall (1978) and Chamberlain (1980). These updating formulae require only inversion of K-dimensional matrices.

and will possess the properties of standard OLS estimation. It is worth emphasizing that because of the nonlinearity of the first-stage estimator, Amemiya's (1978) equivalence results for one- and two-stage estimation procedures are not applicable. Note further the important fact that we may relax the group error normality assumption with no substantive effect on the asymptotic bias [though $(\hat{d} - d_0)$ will no longer be asymptotically normally distributed].

In practice, the errors will be nonspherical, with covariances depending upon the variability of parameter estimates from the first-stage ML model. The variance matrix for w is approximated in finite samples by $\Omega = \sigma_u^2 I_J + V_{dd}$, where V_{dd} is the portion of the first-stage variance matrix corresponding to the dummy variables. This expression provides the intuitive result that using the estimated dummy variables in the second stage adds the estimator variance for the $\hat{d}$ from the first-stage to the variance matrix for the residual u. Let $\Lambda_0 = (\beta'_0, \gamma'_0)$ and define the two-stage estimator $\hat{\Lambda}' = (\hat{\beta}', \tilde{\gamma}')$, where the $\hat{\beta}$ is derived from the ML estimator of the fixed effects model and where $\tilde{\gamma}$ are estimates derived from GLS regression of the $\hat{d}$ on the Z; $\tilde{\gamma} = (Z'\hat{\Omega}^{-1}Z)^{-1}Z'\hat{\Omega}^{-1}\hat{d}$ where $\hat{\Omega}$ is an estimate of the second-stage residual variance matrix (see Appendix Section A.1). Given $V_{d\beta}$, the covariance matrix for the first-stage estimators $\hat{d}$ and $\hat{\beta}$, define the matrices $\Sigma_{\gamma\gamma} = (Z'\Omega^{-1}Z)^{-1}$ and $\Sigma_{\gamma\beta} = -(Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}V_{d\beta}$. It follows (from Pierce) that $\hat{\Lambda} - \Lambda_0$ has approximate variance matrix Σ given by

$$\Sigma = \begin{pmatrix} V_{\beta\beta} & \Sigma_{\beta\gamma} \\ \Sigma_{\gamma\beta} & \Sigma_{\gamma\gamma} \end{pmatrix}. \tag{2}$$

For fixed J, V_{dd} and $\Sigma_{\beta\gamma}$ are o_p (1) so that $\tilde{\gamma} - \gamma_0 \overset{a}{\sim} \mathrm{N}(0, \sigma_u^2(Z'Z)^{-1})$ as $N \to \infty$. Note that whether or not the group error u is normally distributed, the second-stage GLS estimator is not in general normally distributed in finite samples since it contains linear combinations of $\hat{d} - d_0$ which may be far from normally distributed for small N.[10]

2.2.3. Large numbers of groups

It is important to emphasize that the asymptotic unbiasedness outlined above and normality results for the two-stage estimation procedure outlined above require only that the group sample sizes approach infinity (for normally distributed group errors). If we also allow the number of groups J to increase with the sample size (such that N/J increases as well), it is possible to demonstrate

[10] It is worth noting by way of comparison that with J fixed, neither the pooled probit, random effects probit, Chamberlain's two-stage estimator, nor Avery, Hansen, and Hotz's approach are consistent and asymptotically normal, nor are they uniformly asymptotically unbiased.

consistency and asymptotic normality for the two-stage estimator of γ for alternative group error distributions. The only theoretical difficulty lies in establishing consistency for the first-stage ML estimates as both J and N/J approach infinity. Previous authors argue heuristically that asymptotic results follow so long as N/J increases (for example, see Heckman, 1981a, or Hsiao, 1986), but the results do not follow from standard theorems (e.g., Wald, 1949) since the parameter space is of variable (increasing) dimension.

The most straightforward way around this difficulty is to stratify the data by group (in a 'reversal' of the first-stage of Chamberlain's approach) and to employ standard $\sqrt{N/J}$ asymptotic results for each group. If efficiency is a concern, the cross-group restrictions can be imposed on the β via minimum distance estimation, or by using the $\hat{d}_j$ as plug-in estimates in ML estimation. While theoretically sound, in practice the stratified approach is likely to be more cumbersome than ML over a large parameter space since it requires estimation of a large number of models. Alternatively, it should be possible to use results from the extensive literature on semi-parametric estimation (see, for example, Wong and Severini, 1991; Ritov, 1991) to establish results for this problem, but this degree of generality strikes us as overkill. Recently, Sueyoshi (1992) has extended and adapted Portnoy's (1984, 1985, 1987) asymptotic results for increasing dimensional linear m-estimators to consider nonlinear regression models with group indicators, and has demonstrated that (in addition to standard assumptions) $(J \log J)/N \to 0$ is sufficient for consistency.[11]

Given consistency and $\sqrt{N/J}$ normality of the first-stage estimates using the stratified model, asymptotic normality for the estimated γ follows immediately from Randles and Pierce and standard proofs. Since the term involving $(\hat{d} - d)$ is $O_p(\sqrt{J/N})$, provided that J grows slowly enough, we may safely ignore it in the second-stage $\sqrt{J}$ asymptotics. Then if $\operatorname{plim}\,(Z'\Omega^{-1}Z)/J = \sigma_u^{-2}\operatorname{plim}\,(Z'Z)/J = \sigma_u^{-2}Q$, where Q is a finite, positive definite matrix, and $\hat{\Omega} \xrightarrow{p} \sigma_u^2 I_J$ as $J \to \infty$, $\operatorname{plim}\,\tilde{\gamma} = \gamma_0$ and $\sqrt{J}(\tilde{\gamma} - \gamma_0) \xrightarrow{d} \mathrm{N}(0,\ \sigma_u^2 Q^{-1})$ as N/J and J approach infinity.

We emphasize that our two-stage approach does not require a distributional choice for u, only the orthogonality conditions. In contrast to the alternative approaches which all require correct specification of the distributions of both the individual effects ε and the group effects u (and generally require normality of u for computational tractability), our procedure requires only that the distribution of ε be correctly specified and that $\hat{d}$ be a consistent estimator. We investigate the practical importance of this robustness in the simulation work below.

[11] Portnoy (1984) establishes the rate $(J \log J)/N$ for pure ANOVA models. See also Drost (1988) for results from a related literature.

3. Simulation results

3.1. Alternative estimators

In this section, we present results for a small, 100 replication, Monte Carlo simulation designed to assess the performance of the various estimators for the group effects probit model. In addition to the pooled probit and Heckman and Willis random effects specification, we initially consider four two-stage estimation techniques. First, we estimate the second-stage model using ordinary least squares. Under the assumptions of our data-generating process, this specification should asymptotically yield the best linear unbiased estimator for γ since V_{dd} is $o_p(1)$. Next, we estimate the model performing a simple heteroskedasticity correction to account for the variability of the first-stage estimates of d, using the asymptotic standard errors for the first-stage fixed effects estimator and the estimator for the variance of the σ_u^2 outlined by Borjas (1987) (see Appendix Section A.1). This estimator accounts only for the unequal variances associated with the estimates of d_j, but ignores the off-diagonal covariance terms in V_{dd}. Finally, we consider two full GLS specifications which incorporate the off-diagonal terms from V_{dd} into the estimate of Ω: one estimate uses the ratio of the estimates from the pooled and the fixed effects specifications as an estimator for σ_u^2 (GLS 1), and the second uses the Borjas estimator for the σ_u^2 (GLS 2).

It is worth noting here that in the simulations reported below, the relative variance of u to the total error variance is chosen to be 0.20 ($\sigma_u^2 = 25$, $\sigma_\varepsilon^2 = 100$). Clearly, the relative performance of the two-stage estimator will improve with greater degrees of group heterogeneity, and it is encouraging to note that the results presented below are obtained for reasonably small (in relative terms) group effects. Further details of the design of the simulation and of further computational considerations are presented in the Appendix.

We begin with a simulation where the individual and group errors are both normally distributed. Table 1 reports summary statistics of simulations for the coefficients of the individual-level X variables. The most striking result is the poor performance of the Heckman–Willis random effects probit estimator. On average, the estimated coefficients are off by about 11 percent. More importantly, despite the fact that the Heckman–Willis approach is the full-information ML estimator, the t-test nulls are rejected well in excess of the nominal size of the test. We attribute this severe bias to some combination of the computational difficulties associated with the Heckman–Willis approach as outlined above.

In contrast, both the pooled and fixed effect probit estimators yield estimated coefficients that are quite close to the true parameter values. The bias for these estimators is generally on the order of 1 to 2 percent of the parameter value. Note, however, that the precision of hypothesis testing is improved substantially when using the fixed effect model. The size of the t-tests is substantially in excess of the nominal size of 5 percent in the estimates obtained from the pooled probit

Table 1
Summary of Monte Carlo results for coefficients of individual level variables (100 replications, 5000 observations, 50 groups)

Specification	True parameter	$\beta_1 = 1.75$	$\beta_2 = 0.5$	$\beta_3 = -1.0$	$\beta_4 = 1.5$
Pooled probit	Bias	0.018	−0.026	−0.010	0.020
	Std. deviation	0.061	0.091	0.058	0.074
	Minimum	1.641	0.206	−1.144	1.326
	Median	1.766	0.473	−1.006	1.527
	Maximum	1.904	0.696	−0.828	1.692
	Size (5% nominal)	0.17	0.07	0.11	0.10
Fixed effects probit	Bias	0.032	−0.013	−0.017	0.033
	Std. deviation	0.045	0.080	0.054	0.067
	Minimum	1.646	0.287	−1.132	1.337
	Median	1.788	0.492	−1.018	1.536
	Maximum	1.869	0.675	−0.852	1.685
	Size (5% nominal)	0.05	0.04	0.10	0.08
Random effects probit (Heckman–Willis)	Bias	−0.199	−0.077	0.115	−0.167
	Std. deviation	0.061	0.072	0.053	0.068
	Minimum	1.392	0.253	−1.017	1.179
	Median	1.551	0.424	−0.881	1.331
	Maximum	1.692	0.598	−0.704	1.489
	Size (5% nominal)	0.95	0.14	0.63	0.73

model. This is not surprising in view of the result from pooled linear models that the standard errors of the group level variables are substantially underestimated (see Moulton, 1986).[12] Furthermore, despite the generally higher biases for the fixed effects model, it is superior to the pooled specification in terms of a mean-square error criterion.

Table 2 reports similar summary statistics for the intercept and for the coefficients of the group level variables. As before, the Heckman–Willis estimator has the poorest performance: the estimated intercept is 13 percent below its true value, and the estimated coefficients of the group level variables are overestimated by an average of 15 percent. Furthermore, while standard tests for the $\hat{\gamma}$ perform somewhat better than for the $\hat{\beta}$, the specification still rejects the null hypothesis eight times more often than the nominal size.

[12] The pooled probit estimated standard errors for the β_1, β_3, and β_4 coefficients are, on average, 21, 9, and 12 percent too low. In contrast, the fixed effects probit coefficients are 9 percent too high, and 5 and 4 percent too low respectively. For β_2, the pooled probit actually does better than the fixed effects estimator, overstating the standard error by 0.3 percent in contrast to the 8 percent for the fixed effects specification. The actual sizes of the asymptotic tests for β_2 are roughly comparable, with the pooled probit rejecting too frequently.

Table 2
Summary of Monte Carlo results for coefficients of group level variables (100 replications, 5000 observations, 50 groups)

True parameters	$\beta_0 = 1.0$		$\gamma_1 = -2.0$		$\gamma_2 = -1.25$	
	Pooled probit			Two-stage OLS		
	β_0	γ_1	γ_2	β_0	γ_1	γ_2
Bias	−0.031	0.021	0.046	0.028	−0.018	0.021
Std. deviation	0.820	0.352	0.424	0.766	0.350	0.420
Minimum	−1.058	−2.837	−2.287	−0.762	−2.838	−2.296
Median	0.897	−1.973	−1.150	0.990	−1.981	−1.200
Maximum	3.275	−1.061	−0.374	3.098	−1.057	−0.307
Size (5% nominal)	0.47	0.50	0.60	0.28	0.32	0.32
	Heteroskedasticity			GLS method 1		
	β_0	γ_1	γ_2	β_0	γ_1	γ_2
Bias	0.005	0.006	0.040	−0.005	0.033	0.057
Std. deviation	0.759	0.347	0.411	0.746	0.342	0.406
Minimum	−0.823	−2.812	−2.227	−0.819	−2.784	−2.201
Median	0.986	−1.959	−1.184	0.974	−1.925	−1.155
Maximum	3.097	−1.052	−0.311	3.050	−1.038	−0.307
Size (5% nominal)	0.06	0.06	0.04	0.04	0.03	0.02
	GLS method 2			Heckman-Willis random effects probit		
	β_0	γ_1	γ_2	β_0	γ_1	γ_2
Bias	−0.009	0.037	0.058	−0.131	0.256	0.221
Std. deviation	0.747	0.341	0.404	0.922	0.369	0.500
Minimum	−0.807	−2.769	−2.195	−1.545	−2.537	−2.298
Median	0.972	−1.928	−1.166	0.819	−1.743	−1.019
Maximum	3.051	−1.035	−0.307	2.876	−0.939	0.245
Size (5% nominal)	0.05	0.05	0.02	0.41	0.44	0.46

It is interesting to note that in terms of bias the pooled probit estimator performs about as well as any of the various two-stage estimators reported in Table 2. Regardless of the estimator used, the estimated intercept and coefficients of the group level variables are off by about 2 to 3 percent. However, the pooled probit estimator again performs poorly in terms of hypothesis testing, with conventional asymptotic t-tests generating sizes of about 50 percent for a nominal 5 percent value. This poor performance reflects the fact that the estimated standard errors are, on average, one-third of the Monte Carlo

Table 3
Summary of Monte Carlo results for extreme value group errors (100 replications, 5000 observations, 50 groups)

	True	Pooled probit			Two-stage GLS		
		Bias	Std. dev.	Size[a]	Bias	Std. dev.	Size
β	1.0	0.537	0.975	0.25	0.384	0.885	0.15
	1.75	0.207	0.069	0.98	0.031	0.046	0.08
	0.5	0.041	0.093	0.08	−0.011	0.084	0.03
	−1.0	−0.120	0.063	0.59	−0.020	0.055	0.07
	1.5	0.182	0.085	0.77	0.031	0.068	0.07
γ	−2.0	−0.248	0.139	0.48	−0.026	0.106	0.00
	−1.25	−0.145	0.187	0.22	−0.004	0.165	0.01

[a] 5% nominal size.

standard errors. The sizes of the tests drop significantly if any of the two-step estimators are used. In particular, the sizes decline to about 30 percent if the second stage is estimated using ordinary least squares, and decline to 5 percent or less if the second stage is estimated using generalized least squares.[13] Note that neither the extent of the bias nor the size of the test are particularly responsive to the type of variance correction made in the second stage. Even a simple heteroskedasticity correction provides estimators that are nearly as good as those using the additional information provided by the nonspherical covariance matrix.[14]

Table 3 reports selected results for the simulation in the case where the group equation does not have normal errors. Not surprisingly, the pooled probit estimates perform quite poorly in terms of bias, with the majority of the bias components comprising close to 90 percent of the root mean-square error. The size computations for the model are correspondingly poor, with, for example, rejections for the β_3 null occurring over half of the time at a 5 percent nominal level. In contrast, the two-stage GLS estimator performs quite well. Ignoring the constant term, estimates of the coefficients are nearly unbiased, and the size computations, while not quite as well-behaved as when the u are normally distributed, are still quite close to the nominal level. If anything, the tests for the γ are too conservative. We attempted to estimate the corresponding Heckman–Willis models but had difficulty attaining convergence (Appendix Section A.2.2).

[13] The OLS estimates understate the standard errors by 45 percent. For comparison purposes, note that the GLS corrected results *overstate* the errors by a few percent.

[14] The latter are, however, better along other dimensions; they are slightly more efficient, with lower standard deviations and tighter ranges than simple heteroskedasticity corrected estimates.

3.2. Group sizes and numbers of groups

Lastly, to provide practical guidance regarding the application of our two-stage GLS approach to data, we consider the finite-sample behavior of the estimator for alternative group sizes and numbers of groups. We estimated 100 replications of balanced designs in which the number of observations per group ranges from 10 to 100 and the number of groups ranges from 10 to 50 (100 to 5000 total observations). Table 4 reports biases and standard errors for the Monte Carlo simulation, as well as the relative importance of the bias component in the mean-square error (MSE). Lastly, for comparison purposes, we compute the empirical rejection frequencies under the assumption that our large-sample normality result is applicable.

There are several results of note. First, in contrast to Heckman's (1981a) influential simulations suggesting that 8 observations per dummy variable may be sufficient for the fixed effects probit estimator to perform well, we find that both the β_3 and the γ_1 bias components are quite high for 10 observations per group. Even for 50 groups (500 total observations), the bias for β_3 is roughly one-fourth of the parameter size, and the squared bias is half of the total MSE. Similar results are obtained for γ_1. The situation improves substantially once we have 25 observations per group, with the bias components falling to under 10 percent for β_3 and γ_1.

Second, provided that we have 25 or more observations per group, the performance of the fixed effects estimator for β_1 appears to depend primarily upon the total sample size, with weak preference given to increasing the number of observations per group relative to the number of groups. For example, the estimator for 25 groups with 100 observations per group is marginally better than 50 groups and 50 observations per group; 1000 observations from 10 groups appears to be slightly preferable to 1100 observations from 25 groups (50 per group).

The behavior of the second-stage estimator for γ_1 is poor when there are only 10 groups, even for large and increasing within group sizes. For example, increasing the total sample size from 250 to 1000 by fixing the number of groups at 10 and increasing group sizes from 25 to 100 observations per group yields only a slight improvement in the standard deviation of the estimator (0.886 to 0.781) and little improvement in the bias component (-0.191 to -0.157). This last result may be compared with the substantive improvement observed in moving from 250 to 625 observations by holding group size constant at 25 and increasing the number of groups from 10 to 25.

Overall, our results suggest that having a greater number of groups with smaller group sizes may improve the performance of the second-stage estimator (provided that the number of observations per group is 'large enough'). For our simulations, we conclude that while 10 observations per group may be too few, once group sizes exceed 25, increasing the number of groups provides greater

Table 4
Monte Carlo results for selected parameters; two-stage models estimated with different group sizes and numbers of groups (GLS method 2).

Obs. per group	Number of groups	$\beta_3 = -1.0$				$\gamma_1 = -2.0$			
		Bias	Std. dev.	Bias^2/MSE	Size[a]	Bias	Std. dev.	Bias^2/MSE	Size
10	10	−0.491	0.785	0.28	0.09	−0.035	1.495	0.00	0.05
	25	−0.286	0.414	0.32	0.15	0.241	0.493	0.19	0.07
	50	−0.258	0.265	0.49	0.22	0.295	0.375	0.38	0.09
25	10	−0.110	0.284	0.13	0.06	−0.191	0.886	0.04	0.10
	25	−0.076	0.162	0.18	0.10	0.039	0.397	0.01	0.03
	50	−0.067	0.113	0.26	0.07	0.039	0.312	0.02	0.03
50	10	−0.047	0.185	0.06	0.07	−0.137	0.792	0.03	0.10
	25	−0.042	0.117	0.11	0.08	−0.008	0.375	0.00	0.03
	50	−0.035	0.080	0.16	0.09	−0.016	0.280	0.00	0.04
100	10	−0.016	0.119	0.02	0.05	−0.157	0.781	0.04	0.10
	25	−0.025	0.076	0.10	0.06	−0.022	0.368	0.00	0.06
	50	−0.021	0.050	0.15	0.05	−0.020	0.271	0.01	0.05

[a] 5% nominal size.

improvement than increasing the observations per group. Put differently, having a large number of noisy estimates of d appears to be preferable to having relatively few, precisely estimated values.

4. Conclusion

In recent years, the group effects, linear regression model has become increasingly popular in applied research both in economics and in other social sciences. In particular, individual outcomes are specified as a function of individual-specific variables as well as a function of group-specific (or 'environmental' variables). The statistical properties of (as well as alternative estimation procedures for) these models, which are a particular formulation of a more general class of random effects models, have been analyzed for the case of continuous (and observable) dependent variables. This paper extends the literature to outline a particular estimation procedure for estimation and inference in probit models with structural group effects.

One key objective guided our analysis: computational tractability. It is well known that random effects models may be particularly difficult to estimate in a nonlinear setting, and that the difficulty grows significantly as the number of groups increases. To avoid these computational problems, we suggest a two-stage estimation procedure. The first-stage estimates a probit with fixed effects; the second stage regresses these estimated fixed effects on the group level variables, correcting for the nonspherical errors. Our approach is analytically and computationally simple so that estimates may be derived using standard econometric software.

Under mild regularity conditions, the second-stage estimates are asymptotically unbiased as the number of observations within groups approach infinity; furthermore, the two-stage estimator is consistent and asymptotically normal as both the number of groups and the number of observations within each group go to infinity (which are precisely the conditions required for consistency by alternative, and much more complex, estimators). More importantly, the results of our Monte Carlo simulation of alternative estimating schemes reveal that the two-stage approach is superior to the one-stage, random effect or pooled probit formulations currently available in the literature.

Appendix

A.1. Variance estimation

The finite-sample correction for the second-stage GLS variance requires estimation of the matrix Ω. The variance matrix V_{dd} may be estimated consistently from the fixed effects ML. σ_u^2 may be estimated in at least two distinct ways.[15]

One approach uses estimates from separate maximum likelihood specifications. The fixed effects estimator for β consistently estimates $\beta_0/\sigma_\varepsilon$, and the probit model on the pooled sample for N/J, J increasing provides consistent estimates $\hat{\beta}_P$ of $\beta_0/(\sigma_u^2+\sigma_\varepsilon^2)^{1/2}$. The ratio of coefficients is given by $\hat{\beta}_F/\hat{\beta}_P = (\sigma_u^2+\sigma_\varepsilon^2)^{1/2}/\sigma_\varepsilon$. Then an estimator for the variance is given by $\hat{\sigma}_{\bar{u}} = [(\hat{\beta}_F/\hat{\beta}_P)^2 - 1]^{1/2}$. The ratio of standard errors is overidentified since there are K coefficients in β. This estimator is consistent as the N approaches infinity. Following Heckman's (1981b) suggestion we recommend taking averages over the K estimates.

A second approach follows Borjas (1987). OLS applied to the second-stage regression provides estimates of an error variance which has components from both u/σ_ε and from the estimation error associated with $\hat{d}$. If $\tilde{w}_j$ is the jth estimated residual from the second-stage OLS regression, then the estimate of the residual variance is $\tilde{\sigma}^2 = \sum_{j=1}^{J} \tilde{w}_j^2/(J-M)$. Expanding the definition of w, $\tilde{\sigma}^2 = 1/(J-M)\sum_{j=1}^{J}[\tilde{u}_j^2 + (\hat{d}_j - d_j)^2 + 2\tilde{u}_j(\hat{d}_j - d_j)]$, where $\tilde{u}$ are estimates of the latent residuals. Since the latter term is approximately zero from the independence of ε and u, $\sigma_{\bar{u}}^2$ can be estimated by $\hat{\sigma}_{\bar{u}}^2 = \tilde{\sigma}^2 - \sum_{j=1}^{J}\hat{\sigma}_j^2/(J-M)$ where the $\hat{\sigma}_j$ is the standard error for the jth dummy variable in the first-stage estimation procedure.[16]

A.2. Computational issues

A.2.1. Monte Carlo design

We carry out a variety of estimation procedures for 100 replications of the underlying data-generation process. The data for each replication consist of observations on binary responses and observable data for individuals in groups. We consider both balanced and unbalanced sampling schemes, with group sizes ranging from 10 to 50 (for the balanced case) and from 50 to 150 individual observations (for the unbalanced case). The number of groups also varies.

For concreteness, let us consider the 5000 observation, 50 group simulation (unbalanced). First, we draw 5000 independent observations from a six-dimensional multivariate normal distribution with mean zero and arbitrarily chosen

[15] Note that since the first-stage probit identifies coefficients up to a scale factor, $\hat{d}$ provides estimates of d/σ_ε. The implicit group error variance to be included in Ω is $\sigma_{\bar{u}}^2 = (\sigma_u/\sigma_\varepsilon)^2$.

[16] As is the case for other GLS estimators based upon differences, estimates of $\hat{\sigma}_{\bar{u}}$ are not guaranteed to be positive. Due to the presence of the first-stage variance matrix $\hat{V}_{dd}$, even if the estimates of the variance are negative, it is possible for the estimated Ω to be positive definite. Our experience is that the variance will be positive for a well-specified model and that negative estimated values may therefore be evidence of misspecification.

variance matrix.[17] We assign to each individual a group identifier on the basis of exogenously specified group sizes. Additionally, the first four variables are treated as individual-specific variables so that we assign to each individual the corresponding values for X; the latter two are deemed group variables, with all individuals in a group assigned Z values based upon the data for the first individual in the group.[18] All of the subsequent analysis is carried out conditional on this realization of the observable variables.

Conditional on the observed data, we consider two simulations for the unobserved components: one in which the group errors are normally distributed and a second one in which the group errors are distributed as extreme value variables. For the first simulation, we repeatedly draw sets of 5000 independent errors ε, from a normal distribution with mean $\beta_0 = 1$ and variance 100, and 50 draws of normally distributed u with mean 0 and variance 25, and assign each individual the corresponding individual ε value and the group error u associated with the draw for the group. The four β coefficients are arbitrarily chosen to be (1.75, 0.5, − 1.0, 1.5) and the two corresponding γ are (− 2.0, − 1.25). The coefficients were chosen so that the signal to noise for the total model is roughly 2:1. We use these individual and group errors to form $Y^*_{ij} = X'_{ij}\beta + Z'_j\gamma + u_j + \varepsilon_{ij}$ for each individual, and assign binary outcomes to Y_{ij}, with $Y_{ij} = 1(Y^*_{ij} > 0)$. The second simulation generates the u from the extreme value distribution.[19] Using the ε drawn previously, the remainder of the steps in the above simulation are then repeated for the new data generating process.

A.2.2. Random effects probit

Our code for estimating the random effects model is a reimplementation of the Gauss–Hermite integration algorithms described in Butler and Moffitt.[20] We

---

[17] Details on the exact design of the simulation are available upon request. Independent, pseudo-random normal deviates are generated by successive calls to the IMSL double precision function DRNNOR (IMSL, 1987). The resulting matrix of independent normal deviates is then multiplied by the Cholesky factorization of the appropriate variance matrix to generate the desired covariance structure.

[18] The two-stage procedure for assignment of group characteristics is purely for programming convenience and should not affect our results.

[19] The pseudo-random variables are obtained from repeated calls to the double-precision IMSL routine DRNEXP which yields standard exponential deviates, and then taking natural logarithms. To reduce sampling variability, the ε are reused from the previous analysis; thus, the only elements that differ across the two simulations are the u. Because the standard form for the extreme value has $E(u) = -0.17444$ and $\mathrm{var}(u) = \pi^2/6$, we transform the resulting errors to have mean 0 and variance 25 for more direct comparability to the standard normal described above.

[20] We thank Robert Moffitt for providing us with a copy of their original FORTRAN code which we used as a guide in our programming and which allowed to verify the accuracy of our results. All of our FORTRAN code is available upon request for a nominal handling charge.

experimented with several choices for the number of expansion terms and settled on four as the basis for the results presented in the text.[21] Interestingly enough, larger numbers of expansion terms began to generate numerical difficulties as the Monte Carlo estimation began to experience numerical underflow on a number of replications. This outcome suggests that it is the errors in the numeric approximation to the likelihood function that are allowing us to estimate the model; as we add terms so that our approximation becomes finer, the likelihood contributions for groups approach zero and the model becomes unstable. Any interpretation of the random effects results reported in the text should bear in mind the likely inherent problems with the estimates.

We note also that the Monte Carlo simulation for the Heckman–Willis estimator exhibited considerable numerical difficulties for the extreme value specification. Out of the first twenty models, the first replication did not converge, and a number of subsequent specifications generated double precision, floating point exceptions. We conclude that nonnormal group errors may make the ML random effects estimator more difficult to estimate.

References

Amemiya, T., 1978, A note on a random coefficients model, International Economic Review 19, 793–796.

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

Anderson, E.B., 1970, Asymptotic properties of conditional maximum likelihood estimators, Journal of the Royal Statistical Society B 32, 283–301.

Avery, R.B., L.P. Hansen, and V.J. Hotz, 1983, Multiperiod probit models and orthogonality condition estimation, International Economic Review 24, 21–35.

Blanchflower, D.G. and A.J. Oswald, 1990, The wage curve, Scandinavian Journal of Economics 92, 215–235.

Borjas, G.J., 1987, Self-selection and the earnings of immigrants, American Economic Review 77, 531–553.

Borjas, G.J., 1992, Ethnic capital and intergenerational mobility, Quarterly Journal of Economics 107, 123–150.

Butler, J.S. and R. Moffitt, 1982, A computationally efficient quadrature procedure for the one-factor multinomial probit model, Econometrica 50, 761–764.

Case, A.C. and L.F. Katz, 1991, The company you keep: The effects of family and neighborhood on disadvantaged youths, NBER working paper no. 3705, May.

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

Chamberlain, G., 1984, Panel data, Ch. 22 in: Z. Griliches and M. Intriligator, eds., Handbook of econometrics, Vol. 2 (North-Holland, Amsterdam).

Drost, F.C., 1988, Asymptotics for generalized chi-square goodness of fit tests (Stichting Mathematisch Centrum, Amsterdam).

Hall, B.H., 1978, A general framework for time series-cross section estimation, Annales de l'INSEE 30–31, 177–202.

[21] Butler and Moffitt provide simulation results suggesting that two may be sufficient.

Hausman, J.A. and W.E. Taylor, 1981, Panel data and unobservable individual effects, Econometrica 49, 1377–1398.

Heckman, J.J., 1981a, The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, Ch. 4 in: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Heckman, J.J., 1981b, Statistical models for discrete panel data, Ch. 3 in: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Heckman, J.J. and R.J. Willis, 1975, Estimation of a stochastic model of reproduction: An econometric approach, in: N. Terleckyj, ed., Household production and consumption (Columbia University Press, New York, NY).

Hsiao, C., 1986, Analysis of panel data (Cambridge University Press, Cambridge).

IMSL, 1987, User's manual: Stat/Library, April, Version 1.0 (IMSL, Houston, TX).

Judge, G.G., W. Griffiths, R.C. Hill, H. Lütkepohl, and T.-C. Lee, 1988, The theory and practice of econometrics, 2nd ed. (Wiley, New York, NY).

Maddala, G.S., 1986, Limited dependent variable models using panel data, Journal of Human Resources 22, 307–338.

Moulton, B.R., 1986, Random group effects and the precision of regression estimates, Journal of Econometrics 32, 385–397.

Mundlak, Y., 1978, On the pooling of cross-section and time-series data, Econometrica 46, 69–86.

Neyman, J. and E.L. Scott, 1948, Consistent estimates based on partially consistent observations, Econometrica 16, 1–32.

Pierce, D.A., 1982, The asymptotic effect of substituting estimators for parameters in certain types of statistics, Annals of Statistics 10, 475–478.

Portnoy, S., 1984, Asymptotic behavior of m estimators of p regression parameters when p^2/n is large: I. Consistency, Annals of Statistics 12, 1298–1309.

Portnoy, S., 1985, Asymptotic behavior of m estimators of p regression parameters when p^2/n is large: II. Normal approximation, Annals of Statistics 13, 1403–1417.

Portnoy, S., 1987, A central limit theorem applicable to robust regression estimators, Journal of Multivariate Analysis 22, 24–50.

Randles, R.H., 1982, On the asymptotic normality of statistics with estimated parameters, Annals of Statistics 10, 462–474.

Rauch, J.E., 1992, Productivity benefits from geographic concentration of human capital: Evidence from the cities, Journal of Urban Economics, forthcoming.

Ritov, Y., 1991, Estimating functions in semi-parametric models, Ch. 24 in: V.P. Godambe, ed., Estimating functions (Oxford University Press, New York, NY).

Robinson, P.M., 1982, On the asymptotic properties of estimators of models containing limited dependent variables, Econometrica 50, 27–41.

Sueyoshi, G.T., 1992, On results for weak consistency of regression models with increasing dimension ANOVA designs, Sept. (University of California, San Diego, CA).

Wald, A., 1949, Note on the consistency of the maximum likelihood estimate, Annals of Mathematical Statistics 20, 595–601.

Wong, W.H. and T.A. Severini, 1991, On maximum likelihood estimation in infinite dimensional parameter spaces, Annals of Statistics 19, 603–632.