Borrower: IOD

Lending String: *GYG,UNH,ZGM,WYU,VZP

Patron:

Journal Title: Journal of econometrics.

Volume: 59 **Issue:** 3
Month/Year: 1993 **Pages:** 301-317

Article Author: Guilkey, David K

Article Title: Estimation and testing in the random effects probit model

Imprint: [Amsterdam] Elsevier

ILL Number: 209289029

Call #:

Location: 1st floor

ODYSSEY ENABLED

Charge
Maxcost: 25.00IFM

Shipping Address:
Drake University, Cowles Library ILL
2725 University Avenue
Des Moines, Iowa 50311
United States

Fax: 515-271-3933
Ariel:
Email: cowles-ILL@drake.edu

JECMB6 59(3)213–408
(1993)

Volume 59, No. 3 October 1993 ISSN 0 304-4076

JOURNAL OF Econometrics

CONTENTS

CONTENTS CONTINUED ON BACK COVER

NORTH-HOLLAND

Journal of Econometrics 59 (1993) 301–317. North-Holland

Estimation and testing in the random effects probit model

David K. Guilkey and James L. Murphy
University of North Carolina, Chapel Hill, NC 27599, USA

Received November 1989, final version received June 1992

This paper examines the finite-sample properties of the random effects probit estimator in comparison to the standard probit estimator and the standard probit estimator with a corrected asymptotic covariance matrix. The Monte Carlo experiment considers data-generating processes consistent with longitudinal data and also data from sample surveys. The probit estimator with corrected asymptotic covariance matrix works surprisingly well over a wide range of parametric configurations and is recommended as long as an estimate of the error correlation is not of high importance.

1. Introduction

Error term specifications of the form

$$v_{ti} = \mu_i + \varepsilon_{ti}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T_i, \tag{1}$$

have had a long history in empirical literatures involved in the analysis of survey data sets. This formulation has proven useful in the analysis of both cross-sectional and longitudinal data sets.

In cross-sectional data sets, the subscript i might refer to observations from the ith sampling unit, while t refers to the tth observation from that unit. The error term v_{ti} is decomposed into μ_i, which does not vary within the sampling unit, and ε_{ti}, which varies both within the sampling unit and across units. Sampling unit could refer to clusters in a sampling design or could refer to communities in a contextual analysis along the lines of Mason, Long, and Entwisle (1983). The basic idea is that if μ_i and ε_{ti} are assumed to be mutually independent and if each is an independent, identically distributed random variable with mean zero and variances σ_μ^2 and σ_ε^2, respectively, then (1) is a parsimonious way of modeling positive correlation among observations within sampling units.

Correspondence to: James L. Murphy, Gardner Hall, Department of Economics, University of North Carolina, Chapel Hill, NC 27599-3305, USA.

In longitudinal or panel data sets, the subscripts refer to individual i at time t. The u_i represent unobserved individual characteristics that do not vary with time, and the ε_{ti} represent unobserved individual characteristics that vary with time.

A variety of estimators have been proposed for regression models with continuous dependent variables and error specification (1). See Judge et al. (1985) or Hsiao (1986) for a review. Probably the best-known estimator is due to Fuller and Battese (1973) which is implemented in the SUPERCARP computer program. The finite-sample performance of the various estimators has been compared in Monte Carlo experiments; see Judge et al. (1985) for references. While ordinary least squares (OLS) provides a consistent estimator, the reported standard errors are incorrect. Scott and Holt (1982) document the fact that the OLS point estimator is not much less efficient than the GLS estimator, but the standard errors are very biased so that statistical inferences can be quite misleading.

While the finite-sample performance of estimators with the random effects error specification are now fairly well documented for models with continuous dependent variables, essentially no work has been done for models with limited dependent variables. The purpose of this paper is to examine the finite-sample performance of the random effects probit estimator and compare it to the standard probit estimator. The random effects estimator is potentially a very important estimator due to the large number of analyses that involve survey data sets of either cross-sectional or longitudinal form and the frequency with which the dependent variable is only measured by a dichotomous response. The standard probit estimator is heavily used in many disciplines. This paper will show that this estimator with no correction for error correlation will frequently lead to incorrect results since the coefficient standard errors are very badly biased. The most important result obtained in this paper is that if the standard errors of the coefficients for the standard probit are suitably corrected for autocorrelation, the resulting estimator dominates the uncorrected probit and performs almost as well as the maximum likelihood estimator which is the random effects probit in this case. This result has important implications for the benefits that can be obtained from the use of the probit estimator with corrected standard errors to take into account design effects that result from cluster sampling.

This paper is organized as follows. Section 2 presents the statistical specifications of the competing estimators. Section 3 presents the format and results of a Monte Carlo experiment that examines their finite-sample performance. We conclude in section 4.

2. The estimators

The first use of the random effects probit model was by Heckman and Willis (1976). Heckman (1981) presents a detailed discussion of the estimator and

Maddala (1987) and Hsiao (1986) provide excellent reviews. The model is

$$Y_{ti}^{*} = x_{ti}\beta + \mu_i + \varepsilon_{ti}, \tag{2}$$

where X_{ti} is a $1 \times k$ vector of exogenous variables, β is a $k \times 1$ vector of coefficients, $\mu_i \sim \mathrm{IN}(0, \sigma_\mu^2)$, $\varepsilon_{ti} \sim \mathrm{IN}(0, \sigma_\varepsilon^2)$, and μ_i and ε_{ti} are mutually independent. Y_{ti}^{*} is an unobserved latent variable. The observed random variable, Y_{ti}, is defined by

$$Y_{ti} = \begin{cases} 0 & \text{if} \quad Y_{ti}^{*} \leq 0, \\ 1 & \text{if} \quad Y_{ti}^{*} > 0. \end{cases}$$

Let $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$, $\rho = \sigma_\mu^2/\sigma^2$, and impose the normalization $\sigma^2 = 1$. Then, if $Y_i = [Y_{1i}, Y_{2i}, \ldots, Y_{Ti}]$ is the observed sequence for individual i, and defining $\tilde{\mu}_i = \mu_i/\sigma_\mu$,

$$P(Y_1) = \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \Phi \left\{ \left[(X_{ti}\beta/\sigma_e) + \tilde{\mu}_i \left(\frac{\rho}{1-\rho} \right)^{1/2} \right] [2Y_{ti} - 1] \right\} \times f(\tilde{\mu}_i)\, \mathrm{d}\tilde{\mu}_i. \tag{3}$$

The likelihood function for the observed sample of Y_{ti}'s is

$$L = \prod_{i=1}^{N} P(Y_i), \tag{4}$$

which can be maximized with respect to $\beta/\sigma_\varepsilon^2$ and ρ, to obtain consistent and asymptotically efficient estimators. This estimator will be referred to as MLE.

A major computational problem is the evaluation of the integral in (3). Butler and Moffit (1982) suggest the use of Hermite integration [see Salzer, Zucker, and Capuano (1952) and Stroud (1974)] along with the BHHH maximization algorithm [Berndt, Hall, Hall, and Hausman (1974)]. We used their method along with a method that used the DFP algorithm in the GQOPT program along with the Hermite integration. A third method of estimation for MLE that was examined involves replacing the Hermite integration used to evaluate (3). We make use of the fact that $\tilde{\mu}_i$ is distributed IN(0, 1) to make random draws from this distribution. For each draw, (3) is evaluated. Unconditional estimates are then obtained by averaging. We felt that this last method may be of some interest since it is similar to the method of simulated moments [e.g., McFadden (1989)] in the sense that it involves averaging over random draws to obtain joint probabilities. It is also somewhat flexible since it is easy to make alternative distributional assumptions about μ. All that would be involved is drawing random numbers from an alternative distribution and then averaging. Unlike McFadden's estimator, this estimator is not consistent unless the number of draws used in the averaging is allowed to go to infinity [see Lerman and Manski (1981)].

Analytic first derivatives were used in all the maximization routines. An approximation to the asymptotic covariance matrix was obtained by using the first derivatives and subroutine OPTMOV in GQOPT. Standard probit estimates were used as starting values for β and the initial value for ρ was set at 0.5.

The standard probit estimator amounts to setting $\mu_i = 0$ for all i in eq. (2). The likelihood function simplifies to

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T_i} \Phi(X_{ti}\beta)^{Y_{ti}} [1 - \phi(X_{ti}\beta)]^{(1-Y_{ti})}. \tag{5}$$

The first derivatives of log $L = \mathscr{L}$ are

$$\frac{\partial \mathscr{L}_{ti}}{\partial \beta} = \frac{[Y_{ti} - \phi(X_{ti}\beta)]\phi(X_{ti}\beta)X'_{ti}}{\phi(X_{ti}\beta)[1 - \phi(X_{ti}\beta)]}, \tag{6}$$

where ϕ is the standard normal density function. The information matrix is

$$I(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\phi^2(X_{ti}\beta)\, X'_{ti}X_{ti}}{\phi[x_{ti}\beta)[1 - \phi(x_{ti}\beta)]}. \tag{7}$$

Eq. (5) was maximized by the method of scoring using (6) and (7). This method was found to be the best method in a recent simulation by Griffiths, Hill, and Pope (1987), both in terms of estimation and in terms of tests of hypotheses where (7) is used to compute standard errors. This estimator, which was shown to be consistent by Robinson (1982), will be referred to as *PROBIT*.

The final estimator uses the *PROBIT* point estimates of β, say $\hat{\beta}$, and a covariance matrix that is corrected for the correlated error structure along the lines of Avery and Hotz (1985) [see also Binder (1981), Fuller (1975), and Gallant and White (1988)]. Let

$$M_i(\beta) = \sum_{t=1}^{T_i} \frac{\partial \mathscr{L}_{ti}}{\partial \beta} \quad \text{and} \quad M(\beta) = \sum_{i=1}^{N} M_i(\beta)\, M_i(\beta)'.$$

The estimated corrected covariance matrix is

$$C(\hat{\beta}) = I(\hat{\beta})^{-1} M(\hat{\beta}) I(\hat{\beta})^{-1}.$$

3. Monte Carlo experiment

Because the finite-sample properties of the random effects probit estimator are unknwon, we performed the simulations described below.

3.1. Format of the experiments

Eq. (2) is simplified to

$$Y_{ti}^* = \beta_1 + \beta_2 x_{ti} + \mu_i + \varepsilon_{ti}, \tag{8}$$

for the Monte Carlo experiment where all terms are scalars. T_i was not allowed to vary within an experiment, so its subscript is dropped. The X_{ti}'s, μ_i's, and ε_{ti}'s were generated IN(0, 1) using the GNML pseudo-random number generator in IMSL. After X_{ti}, μ_i, and ε_{ti} were drawn, Y_{ti}^* was constructed using (8) and a particular choice of β_1 and β_2. The Y_{ti}'s were then calculated as follows:

$$Y_{ti} = \begin{cases} 0 & \text{if} \quad Y_{ti}^* \le 0, \\ 1 & \text{if} \quad Y_{ti}^* > 0. \end{cases}$$

New samples of X's, μ's, and ε's were drawn on each replication of an experiment. Depending upon the total sample size, we used either 500 or 250 replications in the experiments. This number of replications (*NREP*) was then included as a control variable in the response surface analysis that is discussed in section 3.3.

In the continuous dependent variable case, the design effect or bias in OLS standard errors is directly related to ρ and T. In the standard probit model, the split in the dependent variable is an important determinant of estimator variance. It is easy to manipulate the split in the dependent variable by changing the value of the constant (β_1). We considered these three parameters (β_1, ρ, and T) the parameters of primary interest in the simulations. We are most interested in the statistical properties of the alternative estimators for β_2. We tried two values, 0 and 1, so that we could examine the test statistics both in terms of size and power.

The three values of ρ (0.25, 0.5, 0.75), three values of β_1 (0, 0.5, 1.0), three values of T (2, 5, 10), and two values of β_2 (0, 1) with a sample size of $N = 100$ generated the basic set of experimental results. Additional runs were generated to gain further insights into questions raised by this first set of experiments. A sample size of 250 ($N = 250$) was tried when it became clear that the MLE estimator was not performing well for some parametric configurations, and values of $T = 3$ and 4 were tried when we saw that the simple probit estimator dominated the MLE estimator for $T = 2$ and $\rho = 0.25$. When the dominance of *CPROBIT* over the other estimators became clear in the initial set of runs, we performed additional experiments to see how small the sample size could be reduced and maintain good performance from this estimator.

An additional set of experiments was done to examine the sensitivity of the root mean squared error of the MLE estimator to the number of Hermite points used to approximate the integral in eq. (3). We tried as few as three points to as many as sixteen points. Substantial improvement was noted until eight points was reached. There was little gain in going beyond eight points except at the extremes of the parametric configurations that were used in the experiments. If $T=10$ and $\rho=0.75$, sixteen points were needed in order to obtain stable results. For the sake of consistency, all of the results that we report for the MLE in any particular table are based on the same number of Hermite points.

All of the reported experiments for the *CPROBIT* estimator are based on X's that were generated from a normal distribution. Additional experiments that are not reported used a uniform distribution ($-0.5<X<0.5$), a chi-squared distribution with one degree of freedom, and a log normal distribution based on a normal with mean zero and variance one. In all cases, the basic results are unchanged.

3.2. Results of the experiments

A large number of summary measures were calculated. If θ represents the true value of a parameter and $\hat{\theta}_i$ ($i=1,\ldots,NREP$) represents the estimated value for replication i, then the statistics are:

1. Bias ($BIAS$) $= \dfrac{1}{NREP}\sum_{i=1}^{NREP}(\hat{\theta}_i-\theta)$.

2. Root mean squared error ($RMSE$) $= \dfrac{1}{NREP}\left[\sum_{i=1}^{NREP}(\hat{\theta}_i-\theta)^2\right]^{1/2}$.

3. Mean absloute deviation (MAD) $= \dfrac{1}{NREP}\sum_{i=1}^{NREP}\left|\hat{\theta}_i-\theta\right|$.

4. Standard deviation (SD) $= \dfrac{1}{(NREP-1)}\sum_{i=1}^{NREP}\left[(\hat{\theta}_i-\bar{\hat{\theta}})^2\right]^{1/2}$.

5. The average estimated standard error of $\hat{\theta}_i(ASD) = \dfrac{1}{NREP}\sum_{i=1}^{NREP} SE(\hat{\theta}_i)$,

   where $SE(\hat{\theta}_i)$ is the standard error of $\hat{\theta}_i$ estimated at each replication of the experiment using the appropriate covariance matrix for either *PROBIT*, *CPROBIT*, or *MLE* described above.

6. The mean and variance of $t_i = \dfrac{\hat{\theta}_i - \theta}{SE(\hat{\theta}_i)}$.

7. The proportion of rejections (*REJECTS*) at the 5% level of significance of the null hypothesis that $\theta = 0$ against a two-sided alternative.

8. A chi-squared (*CHI-SQ*) goodness-of-fit test for normality using 20 equi-probable cells and the t_i's generated in item 6.

The results for the BHHH algorithm and the DFP algorithm were essentially the same for $\rho = 0.5$ except the BHHH was more expensive. For $\rho = 0.25$ and $\rho = 0.75$, the BHHH algorithm was unreliable in the sense that it frequently forced ρ to either zero or one. Since the results for BHHH were the same as for DFP when BHHH worked, we do not report separate results for this method. A clear implication of our work is that the DFP algorithm is the preferred method for this model.

The choice between *MLE* with Hermite integration and *MLE* with simulated moments requires additional discussion. We computed the ratio of the *RMSE* of the simulation estimator to the *RMSE* of the estimator that uses Hermite integration. The *MLE* estimator with ten Hermite points is compared to the simulation estimator with ten, fifteen, twenty, and fifty draws. With $N = 100$, $T = 5$, $\rho = 0.5$, and $\beta_2 = 0$, we obtain the following ratios:

	Draws			
Intercept	10	15	20	50
$\beta_1 = 0.0$	1.0607	1.0572	1.0698	0.9830
$\beta_1 = 0.5$	1.0705	1.0298	1.0755	0.9600
$\beta_1 = 1.0$	1.1013	1.0261	1.1097	0.9600

Values of the ratio of *RMSE*'s greater than one indicate the superiority of Hermite integration. In all cases, for fewer than fifty draws the *MLE* estimator with Hermite integration is superior to the simulation estimator. As expected, the simulation estimator tends to improve, but not uniformly, as the number of draws used to compute the average increases. If fifty draws are used for the simulation estimator, it has smaller *RMSE* than the Hermite estimator, and so it is clear that the approximation error can be made arbitrarily small. Since the computational cost of one Hermite point is approximately the same as the cost of one draw for the simulation estimator, the Hermite estimator is clearly preferred on computational grounds. However, the good performance of the simulation estimator is of interest because of its simplicity and the fact that it is straightforward to examine the sensitivity of one's results to the assumed distribution of μ_i simply by changing the distribution from which the random numbers are drawn.

Table 1 presents the results for the basic set of experiments with $\beta_2 = 0$. The results for $\beta_2 = 1$ are similar and are not reported in order to save space. We chose to present results for root mean squared error, proportion of rejections, the goodness-of-fit test, and the ratio of the standard deviation to the average estimated standard error. The other statistics are not reported to conserve space and because the statistics reported give a clear picture of the finite-sample behavior of the two estimators. We now discuss each in turn.

PROBIT and *MLE* differed little in terms of *RMSE*; *CPROBIT* and *PROBIT*, of course, have the same *RMSE*'s. Both estimators are consistent and the unreported bias measure revealed almost no difference on average between the true value of β_2 and its estimated value. Thus, the *RMSE*'s are essentially standard deviations. It is clear that very little efficiency is gained in going from the simple probit estimator to the random effects estimator. Remember that sixteen Hermite points were used for the *MLE*. If ten points are used, similar results are obtained except when $T = 10$ and $\rho = 0.75$. At these extreme values, the *MLE* did substantially worse than the simple probit estimator for this smaller number of points. This degradation of the performance of the *MLE* estimator based on ten Hermite points as T and ρ increase is also noted in the response surface results reported below. A clear implication of these results is that one needs to choose the number of points used in the approximation with care as these two parameters increase. We will return to this point later in the paper. The unreported mean absolute deviations revealed no additional insights.

While *PROBIT* provides good point estimates of the coefficients, the other statistics reveal that misleading results will be obtained if the reported standard errors are used for hypothesis testing. The problem with the standard probit estimator is clearly illuminated in an examination of the *SD/ASD* columns of table 1 and table 2. *SD* is an estimate of the standard error of $\hat{\beta}_2$ obtained through replication in the Monte Carlo experiment, while *ASD* is calculated by averaging the standard error estimates obtained in each replication from the square root of the appropriate diagonal element of the estimated information matrix. *SD* will approach the true standard error of the estimator as the number of replications increases, and we use it as a benchmark by which to judge *ASD*. In an empirical application, standard errors would, of course, be calculated using some approximation to the information matrix such as we use to calculate *ASD*. An *SD/ASD* ratio less than one would indicate that the *PROBIT* standard errors are biased upwards, while a ratio greater than one would indicate that the *PROBIT* standard errors are biased downwards.

In table 1, there is no case for *PROBIT* where the ratio is less than one. The excess of the ratio above 1.0 is small for $T = 2$ and $\rho = 0.25$, but grows more

Table 1

Comparison of *PROBIT*, *CPROBIT*, and *MLE*; $N = 100$, $NREP = 500$, $\beta_2 = 0$.

	$\rho = 0.25$				$\rho = 0.5$				$\rho = 0.75$			
	RMSE	*REJECTS*	*CHI-SQ*	*SD/ASD*	*RMSE*	*REJECTS*	*CHI-SQ*	*SD/ASD*	*RMSE*	*REJECTS*	*CHI-SQ*	*SD/ASD*
$T = 2$												
$\beta_1 = 0.0$												
PROBIT	0.0972	0.068	20.2	1.07	0.1034	0.086	27.8	1.14	0.1138	0.122	70.2	1.24
CPROBIT		0.060	29.0	1.02		0.048	12.9	0.97		0.040	19.0	0.99
MLE	0.1005	0.046	56.1	0.39	0.1032	0.044	40.1	0.62	0.1137	0.030	81.9	0.98
$\beta_1 = 0.5$												
PROBIT	0.1031	0.078	11.4	1.08	0.1108	0.082	43.8	1.16	0.1175	0.104	54.7	1.28
CPROBIT		0.056	17.8	0.99		0.048	19.9	0.96		0.052	19.0	0.99
MLE	0.1036	0.050	31.8	0.49	0.1113	0.034	24.1	0.95	0.1178	0.032	54.2	0.98
$\beta_1 = 1.0$												
PROBIT	0.1139	0.058	12.0	1.03	0.1175	0.056	16.2	1.06	0.1326	0.100	58.1	1.19
CPROBIT		0.044	23.5	0.96		0.062	21.0	0.98		0.050	14.6	0.99
MLE	0.1133	0.036	110.2	0.72	0.1166	0.026	31.2	0.92	0.1315	0.026	66.5	0.97
$T = 5$												
$\beta_1 = 0.0$												
PROBIT	0.0730	0.126	65.2	1.29	0.0877	0.212	227.8	1.53	0.1043	0.270	417.2	1.82
CPROBIT		0.048	27.6	0.99		0.054	19.8	0.98		0.064	21.0	1.04
MLE	0.0730	0.060	20.8	1.00	0.0873	0.056	41.0	0.93	0.1025	0.052	23.1	1.01
$\beta_1 = 0.5$												
PROBIT	0.0778	0.128	94.0	1.29	0.0904	0.190	201.8	1.51	0.1045	0.266	377.7	1.74
CPROBIT		0.068	22.7	1.05		0.076	27.7	1.05		0.070	32.3	1.05
MLE	0.0777	0.050	22.7	1.03	0.0896	0.058	18.5	1.00	0.1032	0.042	24.6	0.99
$\beta_1 = 1.0$												
PROBIT	0.0826	0.100	39.7	1.19	0.0994	0.180	161.5	1.43	0.1202	0.254	406.6	1.72
CPROBIT		0.068	27.6	1.05		0.078	17.4	1.06		0.074	26.2	1.06
MLE	0.0830	0.042	36.6	0.84	0.0989	0.054	28.4	1.00	0.1181	0.048	27.9	1.02

D

Table 1 (continued)

	$\rho = 0.25$				$\rho = 0.5$				$\rho = 0.75$			
	RMSE	*REJECTS*	*CHI-SQ*	*SD/ASD*	*RMSE*	*REJECTS*	*CHI-SQ*	*SD/ASD*	*RMSE*	*REJECTS*	*CHI-SQ*	*SD/ASD*
						$T = 10$						
$\beta_1 = 0.0$												
PROBIT	0.0622	0.208	217.7	1.54	0.0795	0.316	581.5	1.96	0.0969	0.402	862.0	2.39
CPROBIT		0.044	32.2	0.97		0.048	24.6	0.98		0.050	20.8	0.99
MLE	0.0618	0.052	30.1	0.88	0.0779	0.044	22.2	0.86	0.0956	0.124	49.2	1.14
$\beta_1 = 0.5$												
PROBIT	0.0648	0.204	255.0	1.52	0.0811	0.304	445.6	1.91	0.0986	0.394	805.4	2.32
CPROBIT		0.060	32.8	1.02		0.056	19.5	1.02		0.056	14.7	1.03
MLE	0.0644	0.044	32.6	1.00	0.0800	0.042	31.2	0.98	0.1055	0.144	63.8	1.20
$\beta_1 = 1.0$												
PROBIT	0.0726	0.182	184.0	1.48	0.0926	0.302	452.2	1.88	0.1119	0.398	781.4	2.27
CPROBIT		0.058	9.44	1.03		0.068	37.4	1.04		0.070	33.2	1.06
MLE	0.0710	0.054	41.5	1.02	0.0884	0.050	31.2	1.00	0.1086	0.126	87.0	1.14

substantial as either of these parameters increases. The downward bias in the *PROBIT* standard error is systematic and has a factor of 2.39 in the worst case. Table 2 gives the most complete picture of this result for $\rho = 0.25$. It is clear that using the reported standard errors that are uncorrected for error correlation would lead to misleading statistical inferences for any T greater than 2.

The results for the maximum likelihood estimator are better but not uniformly better. The *MLE* performs poorly for $T=2$ and $\rho = 0.25$, a finding consistent with results found in the continuous dependent variable case [e.g., Nerlove (1971) and Maddala (1987)]. It then improves as T or ρ increases until $T = 10$ and $\rho = 0.75$ where the performance starts to deteriorate slightly. The *SD/ASD* ratio is typically less than one for small ρ and T, indicating that the reported standard errors are too large. This situation is reversed for large ρ and T.

The additional experiments reported in table 2 were run because of this set of inconsistent results for the *MLE*. Table 2 shows that the performance of the *MLE* with $\rho = 0.25$ improves as T increases, with much of the improvement coming in the move from $T = 2$ to $T = 3$. Our conjecture is that at $T = 2$ there is simply not sufficient information available to the *MLE* to allow it to estimate ρ accurately and, as a result, it clearly is not superior to the simple probit estimator which ignores ρ when it estimates the β's. It is comforting to note that the *MLE* shows considerable improvement as sample size increases from 100 to 250, but even at $N = 250$, it does not necessarily offer improvement in terms of the *SD/ASD* ratio over the simple probit estimator for $T = 2$.

The results for the *SD/ASD* ratio are quite impressive for *CPROBIT*. The ratio is very close to one for all experiments. In 14 of the 27 cases in table 1, the ratio is slightly greater than one. In every case the ratio falls within the interval

Table 2

SD/ASD comparisons; $\rho = 0.25$, $\beta_2 = 0$.

	$N = 100$, $NREP = 500$				$N = 250$, $NREP = 250$			
	T				T			
	2	3	4	5	2	3	4	5
$\beta_1 = 0$								
PROBIT	1.09	1.14	1.25	1.24	0.99	1.16	1.19	1.32
CPROBIT	1.02	1.03	0.96	0.99	1.03	0.95	0.99	1.06
MLE	0.40	0.68	1.04	0.97	0.86	1.01	0.98	0.69
$\beta_1 = 0.5$								
PROBIT	1.05	1.16	1.26	1.30	1.09	1.16	1.25	1.25
CPROBIT	0.99	1.02	1.02	1.05	1.02	0.95	1.01	0.97
MLE	0.47	0.77	1.05	1.03	0.72	1.01	1.04	0.99
$\beta_1 = 1.0$								
PROBIT	0.98	1.12	1.19	1.25	1.10	1.13	1.22	1.20
CPROBIT	0.96	1.04	1.07	1.05	1.06	0.89	0.91	1.02
MLE	0.66	0.71	0.77	1.03	1.04	1.05	1.04	0.97

[0.96, 1.06], indicating that tests based on the *CPROBIT* covariance matrix would be quite reliable. The additional results reported in table 2 also indicate that *CPROBIT* performs quite well, even when compared to *MLE* at $N = 250$.

The results for *RMSE* and the *SD/ASD* ratio accurately reflect the results that we obtained for the mean and variance of the t-statistic (not reported). The mean of the t-statistic was close to zero for all estimators. The true variance of the t-statistic should approach one as sample size increases. This was true for *MLE* and *CPROBIT*, but was not true for *PROBIT*. *PROBIT* typically resulted in a variance substantially greater than one.

The *REJECTS* statistic reported in table 1 tells a similar story. This statistic tabulates the proportion of rejections of the null hypothesis $\beta_2 = 0$ against a two-sided alternative at the 5% level of significance. All experiments were conducted with 500 replications, so with $\beta_2 = 0$ and 5% test, a 95% confidence interval would be approximately 0.05 ± 0.02. The *PROBIT* proportion of rejections lies within the 95% confidence interval only 4 of 27 times, while the *MLE*'s proportion of rejections is inside the interval 22 of 27 times. *PROBIT* tends to perform well for $T = 2$ and $\rho = 0.25$ or 0.5. In all other cases, the frequency with which the true null hypothesis is rejected is extremely high with a maximum of 0.402. The problem with *PROBIT* becomes more severe as either T or ρ increases. The *MLE*, on the other hand, performs very well until both T and ρ are at their highest values: $T = 10$ and $\rho = 0.75$. This poor performance of the *MLE* for the proportion of rejections corresponds to the cases in which the *SD/ASD* ratio becomes large, indicating the problem is related to using estimated standard errors that are too small. The performance of the *CPROBIT* estimator completely dominates the other two estimators for the *REJECTS* statistic. Its proportion of rejections is within the 95% confidence interval 24 of 27 times with the three outliers being very close (0.074, 0.076, and 0.078).

There is little to discuss when $\beta_2 = 1$ (results not in table). In this case, all tests rejected the null hypothesis 100% of the time in almost all cases.

Tables 1 and 3 also present results for the χ^2 test defined above as statistic item 8. It is used to test the null hypothesis that the t-statistics come from a standard normal distribution when the true value of β_2 is used in forming the test statistic. The critical values of the chi-square are 30.1 for a 5% test and 36.2 for a 1% test. We will use the 1% value in our discussion.

The null hypothesis of normality is rejected 22 out of 27 times in table 1 for the *PROBIT*. The implications mirror those for the other statistics in the sense that *PROBIT* tends to perform well for $T = 2$ and $\rho = 0.25$ or 0.5. The *MLE* performs better, but the null hypothesis is still rejected 12 out of 27 times. Good performance is typically noted for intermediate values of T and ρ. Table 3 clearly demonstrates that the poor performance of the *MLE* is related to N. In this table, as N moves from 100 to 250, the number of rejections of the null for the *MLE* goes from 6 of 12 to 0 of 12. *PROBIT*, on the other hand, fails to improve as sample size increases in the N direction. The normality test again

Table 3
Chi-squared test comparisons; $\rho = 0.25$, $\beta_2 = 0$.

	$N = 100$, $NREP = 500$				$N = 250$, $NREP = 250$			
	T				T			
	2	3	4	5	2	3	4	5
$\beta_1 = 0$								
PROBIT	26.5	32.3	61.4	63.9	38.8	38.0	32.4	49.4
CPROBIT	29.0	21.3	19.3	27.6	18.3	11.9	17.0	16.6
MLE	31.5	29.9	37.0	37.5	26.6	15.0	17.5	30.0
$\beta_1 = 0.5$								
PROBIT	15.5	34.0	85.9	111.3	34.8	29.8	46.2	27.2
CPROBIT	17.8	25.9	29.4	22.7	13.4	22.6	23.1	12.7
MLE	36.3	28.9	43.2	35.8	24.6	20.4	7.0	21.4
$\beta_1 = 1.0$								
PROBIT	25.7	24.0	33.6	59.2	25.4	33.4	18.6	34.5
CPROBIT	23.5	17.9	19.4	27.6	16.1	21.3	11.4	23.6
MLE	101.4	22.6	45.2	34.3	21.7	25.4	13.4	23.8

demonstrates the usefulness of the *CPROBIT* estimator. The null hypothesis of normality is only rejected 1 of 27 times in table 1, and there are no rejections at all in table 3.

The impressive performance of *CPROBIT* for $N = 100$ and $N = 250$ led us to an additional set of experiments for smaller sample sizes so that we could determine when one could expect the performance to deteriorate. Table 4 presents results for $N = 50$. The same set of experiments were also performed with $N = 25$ and are not displayed to conserve space. In addition, $T = 25$ was added. It should be noted that we tried to use $T = 25$ in the original set of experiments but were unable to do so because of numerical accuracy problems associated with calculating the joint probability in (3). The table only reports results for $\rho = 0.25$. The results for the larger values of ρ were similar.

The basic conclusion that can be drawn from these additional experiments is that $N = 25$ (results not shown) is not a sufficiently large sample size for the asymptotically valid covariance matrix approximation to be reliable. For example, the normality test was rejected 19 of 24 times, and in 11 of 12 cases for $\beta_2 = 0$ the proportion of rejections is greater than 0.07. Doubling N to 50 improves performance tremendously. Normality is rejected only 3 times, while for $\beta_2 = 0$ the proportion of rejections is within the 95% confidence range 6 of 12 times.

3.3. Response surface analysis

It is well known that it is difficult to generalize Monte Carlo results beyond the specific parametric configurations used in the experiments. We used

Table 4
Additional results for *CPROBIT*; $N = 50$, $\rho = 0.25$.

	$\beta_2 = 0$				$\beta_2 = 1$			
	RMSE	*REJECTS*	*CHI-SQ*	*SD/ASD*	*RMSE*	*REJECTS*	*CHI-SQ*	*SD/ASD*
				$T = 2$				
$\beta_1 = 0.0$	0.1456	0.066	16.2	1.06	0.2497	1.000	40.3	1.13
0.5	0.1524	0.068	34.5	1.08	0.2434	1.000	19.4	1.07
1.0	0.1783	0.082	39.5	1.11	0.2811	1.000	33.5	1.12
				$T = 5$				
$\beta_1 = 0.0$	0.1061	0.074	22.4	1.02	0.1603	1.000	26.8	1.05
0.5	0.1109	0.080	20.6	1.04	0.1546	1.000	17.0	0.99
1.0	0.1232	0.076	22.7	1.05	0.1791	1.000	35.1	1.05
				$T = 10$				
$\beta_1 = 0.0$	0.0880	0.068	15.2	1.00	0.1302	1.000	26.6	1.07
0.5	0.0917	0.068	11.0	1.02	0.1326	1.000	16.6	1.06
1.0	0.1024	0.068	21.4	1.05	0.1467	1.000	40.2	1.08
				$T = 25$				
$\beta_1 = 0.0$	0.0822	0.074	32.1	1.04	0.1062	1.000	12.6	1.02
0.5	0.0843	0.068	22.1	1.05	0.1073	1.000	22.6	1.02
1.0	0.0894	0.082	25.4	1.06	0.1235	1.000	23.4	1.09

response surface analysis along the lines suggested by Hendry (1984) to try to gauge the sensitivity of our results to the parametric configurations used. Table 5 presents the results of the response surface regressions with *RMSE* of β_2 and ρ as the dependent variables. Note that the number of observations in the response surface regression is larger than the number of observations for experiments in tables 1–4. The additional observations were generated in unreported experiments that varied the number of replications in addition to experiments that were added to obtain additional variation in N and β_1. All these variables appear as regressors.

The regression results provide a nice summary of the above discussion, and the results for β_2 and ρ are very similar. We concentrate on the results for β_2 in our discussion. The only insignificant regressors are *ML METHOD* (zero for *PROBIT* and one for the *MLE*) and ρ. Note that both of these variables are significant as components of interaction terms. An increase in the number of replications causes a very slight but significant reduction in *RMSE*. We would have preferred not to have obtained this result since increasing the number of replications in such experiments becomes prohibitively expensive. The fact that the estimated coefficient is so small indicates that this result is of little practical significance.

Table 5
Response surface analysis.

	RMSE of β_2	RMSE of ρ
Constant	0.241 (33.90)	0.384 (22.06)
$\sqrt{T}$	-0.042 (-28.10)	-0.073 (-20.58)
$\sqrt{N}$	-0.008 (-21.25)	-0.010 (-11.11)
ρ	0.003 (0.69)	-0.112 (-11.00)
β_1	0.017 (12.86)	0.019 (6.36)
β_2	0.039 (33.16)	0.011 (3.64)
NREP	-0.00002 (-4.22)	-0.00003 (-2.72)
$\sqrt{T}\times\rho$	0.011 (23.65)	0.12 (11.26)
ML METHOD	-0.003 (-1.61)	
$\sqrt{T}\times\rho\times$ *Method*	0.004 (2.34)	
NPT		-0.0004 (-1.00)
SIM METHOD	0.010 (3.99)	0.009 (1.90)
Sample size	374	193
R^2	0.92	0.81

Two variables, $\sqrt{T}$ and $\sqrt{N}$, cause *RMSE* to decrease when they increase, as expected. β_1 and β_2 have the expected positive effect on *RMSE*. As β_1 increases, the split in the dependent variable gets worse ($\beta_1 = 0$ results in 50% ones, while $\beta_1 = 1$ results in 75% ones) and *RMSE* increases. As β_2 moves from zero to one, the variance in the underlying dependent variable increases and *RMSE* increases.

A number of interaction terms were tried. The two interaction terms reported involve ρ, $\sqrt{T}$, and method. Note that the main effect for method is marginally not significant but negative, indicating a slight reduction in *RMSE* from the use of *MLE*. As the interaction term $\sqrt{T}\times\rho$ increases, *RMSE* deteriorates with a more substantial impact on *MLE* than *PROBIT*. It should be noted that most of the observations used in the response surface analysis involved experiments with the number of hermite points set equal to ten. As stated earlier, it is clear that additional points are needed as T and ρ increase. Because of the large computational cost for the MLE estimator as the number of points is increased, it may be of limited usefulness for large T unless an estimate of ρ is needed.

4. Conclusion

In this paper we examined the finite-sample performance of the standard probit and the random effects probit (MLE) estimators when the error term is autocorrelated due to an error components structure. $PROBIT$ is an extremely important method of estimation in the analysis of survey data sets. We show conclusively that the use of a standard probit in situations in which observations are clusterd either by sample design (T large and N small) or by panel data designs (T small and N large) can lead to very misleading inference.

The asymptotic efficiency of the MLE estimator was not always a good indication of its performance in finite samples. If only two points are available, one may as well use standard probit. MLE is useful in panel designs with intermediate numbers of longitudinal data points, especially if an estimate of ρ is desired. However, as the interaction of the number of points and ρ increases, the performance deteriorates unless the number of Hermite points is increased, which, of course, increases computational cost.

Fortunately, the $CPROBIT$ estimator provides a cost-effective alternative to MLE that performs extremely well for almost all parametric configurations used in this paper. $CPROBIT$ is reliable in both panel data configurations and also cluster sampling configurations.

References

Avery, R. and V.J. Hotz, 1985, HOTZTRAN user's manul (CERA Economic Concultants, Greenwich, CT).

Berndt, E., B. Hall, R. Hall, and J. Hausman, 1974, Estimation and inference in non-linear structural models, Annals of Economic and Social Measurement 3, 653–665.

Binder, D.A., 1981, On the variances of asymptotically normal estimators for complex surveys, Survey Methodology 7, 157–170.

Butler, J.S. and R. Moffit, 1982, A computationally efficient quadrature procedure for the one-factor multinomial probit model, Econometrica 50, 761–765.

Ericsson, N,R., 1987, Monte Carlo methodology and the finite sample properties of statistics for testing nested and non-nested hypotheses, International Finance Discussion Papers no. 317.

Fuller, W.A., 1975, Regression analysis for sample survey, Sankya Series 6, 37, 117–132.

Fuller, W.A. and G. Battese, 1973, Transformations for the estimation of linear models with nested error structure, Journal of the American Statistical Association 68, 626–632.

Gallant, A.R. and H. White, 1988, A unified theory of estimation and inference for nonlinear dynamic models (Basil Blackwell, Oxford).

Griffiths, W.E., R.C. Hill, and P.J. Pope, 1987, Small sample properties of probit model estimators, Journal of the American Statistical Association 82, 929–937.

Heckman, J.J., 1981, Statistical models for discrete panel data, in: C. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).
Heckman, J.J. and R. Willis, 1976, Estimation of a stochastic model of reproduction: An econometric approach, in: N. Terleckyj, ed., Household production and consumption (National Bureau of Economic Research, New York, NY).
Hendry, D.F., 1984, Monte Carlo experimentation in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics (North-Holland, Amsterdam).
Hsiao, C., 1986, Analysis of panel data (Cambridge University Press, New Rochelle, NY).
Judge, G., W.E. Griffiths, R.C. Hill, and T.C. Lee, 1985, The theory and practice of econometrics, 2nd ed., Series in probability and mathematical statistics (Wiley, New York, NY).
Lehman, S.R. and C.F. Manski, 1981, On the use of simulated frequencies to approximate choice probabilities, in: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).
Maddala, G.S., 1987, Limited dependent variable models using panel data, Journal of Human Resources 22, 307–337.
Mason, W.M., G. Long, and B. Entwisle, 1983, Contextual analysis through the multilevel linear model, in: S. Leinhardt, ed., Sociological methodology, 1983–94 (Josey Bass, San Franscisco, CA) 72–103.
McFadden, D., 1989, A method of simulated moments for estimation of discrete response models without numerical integration, Econometrica 59, 995–1026.
Nerlove, M., 1971, Further evidence on the estimation of dynamic economic relations from a time series of cross sections, Econometrica 39, 359–382.
Robinson, P.M., 1982, On the asymptotic properties of estimators of models containing limited dependent variables, Econometrica 50, 27–41.
Salzer, H.E., R. Zucker, and R. Capuano, 1952, Table of the zeros and weight factors of the first twenty hermite polynomials, Journal of Research NBS 48, 111–116.
Scott, A.J. and D. Holt, 1982, The effect of two-stage sampling on ordinary least squares methods, Journal of the American Statistical Association 77, 848–854.
Stroud, A.H., 1974, Numerical quadrature (Springer-Verlag, New York, NY).