GMM Estimation of Panel Probit Models:
Nonparametric Estimation of the Optimal Instruments

Irene Bertschek*) Michael Lechner**)[1]

*) C.O.R.E. and Institut de Statistique, Louvain-la-Neuve
**) University of Mannheim, ZEW, Mannheim, and Harvard University, Cambridge

February 1995

"Comments welcome"

Correspondence to:
Michael Lechner
Center for European Studies
Harvard University
27, Kirkland Street
Cambridge, MA 02138
USA

[1] We gratefully acknowledge financial support by the "Human Capital and Mobility" programme, ERBCHBICT 941032, the Deutsche Forschungsgemeinschaft (DFG) and the SFB373, Humboldt-Universität zu Berlin. We thank Jörg Breitung, Sigbert Klinke, Byeong Park and the participants of econometric seminars at the Universities of Konstanz and Mannheim as well as of the Humboldt-Universität zu Berlin for helpful comments and suggestions on a previous draft of the paper.

Abstract

The generalised method of moments (GMM) is combined with the nonparametric estimation of the instrument matrix to obtain an easily computable estimator for the panel probit model. It is based on the specification of the conditional mean of the binary dependent variable in each period, and therefore high-dimensional numerical integrations are avoided. An extensive Monte Carlo study compares this estimator with other estimators proposed in the literature and shows that it has good small sample properties and that the efficiency loss compared to full information maximum likelihood is quite small. Some of the estimators are applied to two microeconometric examples.

KEY WORDS: panel probit, GMM, conditional mean restrictions, k-nearest neighbor.

1 Introduction

For cross-section estimation of nonlinear models maximum likelihood (ML) is the typically used method. However, using ML on panel data is quite tough due to intertemporal correlations of the error terms. It is necessary to specify the correlation matrix and to evaluate T-variate integrals which is a very computertime-consuming task. One possibility is to use simulation methods in order to obtain approximations of the intergrals (Hajivassiliou, 1993), but these methods are still computer-intensive. Butler and Moffitt (1982) propose a simplification by restricting the covariance to have a one factor random effects structure. Although they introduce an efficient algorithm for estimating ML, there is, to the best of our knowledge no proof available that the suggested estimator remains consistent when the true correlation structure has no one factor representation.

Avery, Hansen and Hotz (1983) suggest to use only marginal moments of the observed dependent variable in each period instead of its joint distribution in all periods. This makes only univariate integration necessary. They base their estimator on the generalised method of moments (GMM) which gives consistent estimates independent of the true covariance structure if appropriate moment conditions are specified. It can be shown that many methods (for instance, pooled probit, Chamberlain's probit) belong to this class of estimators (cf. Breitung and Lechner, 1994, for a comparison). Since these types of GMM estimators do not specify the covariance structure of the errors explicitly, the loss of information leads to a loss of efficiency compared to full information maximum likelihood (FIML). Newey (1993) shows for general non-linear models how to exploit information optimally in order to keep the loss of efficiency as small as possible, by choosing the instruments optimally in combination with the conditional moment restrictions.

Instead of using suboptimal instruments or parametric approximations of the optimal instruments by assuming particular covariance structures as in Breitung and Lechner (1994), we exploit Newey's results that an efficient estimator can be built by the nonparametric method of k-nearest neighbors independently of the true covariance structure. In addition, this estimator is simple to compute and has a good small sample performance as the Monte Carlo results show.

In the next section, notation and assumptions of the model are introduced. Section 3 shows how to obtain an asymptotically efficient estimator given the implied conditional moment restrictions which are at the heart of the analysis. Considerations on nonparametric estimation in this model are contained in section 4. Section 5 presents the design and the results of the Monte Carlo study. The data generating processes are introduced. Different estimators are discussed, such as pooled probit, Chamberlain's sequential probit, GMM using an approximation of the random effects covariance for the optimal instruments and various estimators based on nonparametric estimation of the optimal instruments. Two applications of some of the estimators are presented in section 6. Section 7 contains some concluding remarks. Two appendices give some additional information referring to the Monte Carlo study and to the applications.

2 The Model

Suppose that for T time-periods realizations $(y_i, x_i) = z_i$ from N independent random draws in the joint distribution of the random variables $(Y, X) = Z$ are observed. The $T \times 1$ dimensional vectors Y, y_i denote the dependent variables, and the $T \times K$ dimensional matrices X, x_i represent the explanatory variables. Hence, random variables Z and the data z_i have the dimension $T \times (K+1)$. The dimension T is asssumed to be much smaller than the dimension N. The model is formulated in terms of a latent variable y_{ti}^* which is linearly related to the explanatory variables contained in x_{ti}, the $K \times 1$ dimensional deterministic coefficient vector β and a scalar error term u_{ti}:

$$y_{ti}^* = x_{ti}\beta + u_{ti} \tag{1}$$

The observation rule is:

$$y_{ti} = 1(y_{ti}^* > 0), \quad i = 1, ..., N, \; t = 1, ..., T \tag{2}$$

The indicator function $1(\bullet)$ is one if its argument is true and zero otherwise. The unobserved variables y_{ti}^* and u_{ti} serve merely as a device to rationalise restrictions on certain conditional and joint distributions of Z. The error terms are jointly normally distributed with mean zero and covariance matrix Σ and they are independent of the explanatory variables. The latter assumption excludes the case of fixed effects, if these fixed effects are correlated with regressors. However, the type of correlated random effets probit model suggested by Chamberlain (1980, 1984) is covered in this model by appropriate redefinition of regressors and β. Time-varying coefficients can also be incorporated in this approach by a simple extension of the X-space and appropriate parameter restrictions. To simplify the notation we will not pursue these issues any further. The $(L \times 1\ )$ vector of nonstochastic coefficients γ contains β and parameters of the covariance matrix of the errors Σ. For identification purposes one main-diagonal element of Σ is set to unity. γ^0, β^0, Σ^0 denote the parameter values characterising the true distribution of Z. All the following asymptotic arguments are based on the assumption that the dimension T is fixed whereas the dimension N is increasing.

Within the class of latent variable models based on (1), the binary choice model is the case with minimum information on the dependent variable because only the sign of y_i^* is observed. The implications are at least twofold: First, this poses difficult identification problems for γ and other quantities of interest. These problems are extensively discussed in Manski (1988) and reviewed in Horowitz (1993). Second, all methods which 'work' for this model, could also be applied to the other models in that class when y_i carries more information about y_i^* (see Lechner and Breitung, 1995, for a discussion of the usefulness of these methods for other models).

3 Conditional Moment Restrictions, GMM Estimation and Asymptotic Efficiency

The model presented in the previous section implies the following moment conditions:

$$\begin{aligned} E[M(Z,\theta^0)|X=x_i] &= 0 \\ M(Z,\theta) &= [m_1(Z_1,\theta),...,m_t(Z_t,\theta),...,m_T(Z_T,\theta)]' \\ m_t(Z_t,\theta) &= Y_t - E(Y_t|X=x_i) \end{aligned} \tag{3}$$

where

$$E(Y_t|X=x_i) = \Phi\left(\frac{x_{ti}\beta}{\sigma_t}\right).$$

$\Phi(a)$ denotes the cumulative distribution function of the standardised univariate normal distribution evaluated at a, and θ represents the parameters of interest which are a subset of the parameter vector γ.

The use of these conditional moments for estimation has various advantages: Firstly, their evaluation does not require multidimensional integration, as it is the case with maximum likelihood estimation. Furthermore, they do not depend on the $T(T-1)/2$ off-diagonal elements of Σ , which can be difficult to estimate particularly when T is large. Finally, many popular estimation methods such as the GMM estimators suggested by Avery et al. (1983), the pooled maximum likelihood estimator, and the sequential estimator suggested by Chamberlain (1980, 1984) can be shown to be based on these conditional moments (see Breitung and Lechner, 1994).

Given these conditional moment restrictions the generalised method of moments (GMM) can be used for estimation. Introduced by Hansen (1982) it is based on minimising quadratic forms of functions which are the sample analogs of the respective population moments. Recent important insights in the properties of estimators based on conditional moment restrictions have been obtained by Chamberlain (1987) and Newey (1990). The excellent survey by Newey (1993) summarises these results and elaborates on them. The following exposition borrows from this source.

The unconditional moment restriction based on equation (3) to be used for the estimation is obtained by observing that $M(Z, \theta^0)$ will be uncorrelated with all functions of X, hence:

$$EA(X)M(Z, \theta^0) = 0. \tag{4}$$

$A(X)$ is a $p \times T$ "instrument matrix". An estimate $\hat{\theta}_N$ of θ^0 is obtained by setting a quadratic form of the sample analogs

$$g_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} A(x_i)M(z_i, \theta) \tag{5}$$

close to zero, such that

$$\hat{\theta}_N = \arg \min_{\theta \in \Theta} \; g_N(\theta)' P g_N(\theta). \tag{6}$$

Θ denotes the parameter space, which is part of the $I\!R^K$. Under suitable regularity conditions on g_N and with the positive semi-definite matrix P, $\hat{\theta}_N$ is $\sqrt{N}$-consistent and asymptotically normal (Newey, 1993):

$$\sqrt{N}(\hat{\theta}_N - \theta^0) \xrightarrow{d} N(0, (G'PG)^{-1}G'PVPG(G'PG)^{-1}) \tag{7}$$

where $G = E\left[A(X)\frac{\partial M(Z,\theta^0)}{\partial \theta'}\right]$ and $V = E[A(X)M(X, \theta^0)M(X, \theta^0)'A(X)']$.

The tools to minimise the variance of this estimator are the optimal choice of the instruments $A(X)$ and of the weighting matrix P. As shown by Hansen (1982) in a more general setting, the optimal choice of P is V^{-1} or any consistent estimator of it. Chamberlain (1987) and Newey (1990, 1993) derived the optimal instrument matrix A. Let

$$D(x_i) = E\left[\left.\frac{\partial M(Z,\theta^0)}{\partial\theta'}\right| X = x_i\right],$$

$$\Omega(x_i) = E[M(Z,\theta^0)M(Z,\theta^0)'|X = x_i] \tag{8}$$

The optimal choice for A denoted by A^* equals:

$$A^*(x_i) = C\ D(x_i)'\Omega(x_i)^{-1}, \tag{9}$$

where C is any non-singular $L \times L$ matrix. Note that the column-dimension of A^* equals L, so that the choice of P is irrelevant, since it only represents a weighting of the score. P, $D(x_i)$ and $\Omega(x_i)$ may be substituted by consistent estimates without affecting the asymptotic distribution of the final coefficient estimates. However, finding these consistent estimates may be a formidable task in a complex model, as will be exemplified for the relatively simple panel probit model. To circumvent these problems, Newey (1990, 1993) suggests using nonparametric methods, such as nearest neighbour estimation and series approximations instead, and derives the conditions necessary for these methods to result in consistent and asymptotically efficient estimates of θ (Newey, 1993, theorems 1 and 2). In order to obtain the optimal GMM-estimator based on (3) the following notation is introduced. Let $\theta = (\theta_1', \theta_2')', \theta_1 = \beta/\sigma_1, \theta_2 = (\theta_{22}, ..., \theta_{2T})', \theta_{2t} = \sigma_1/\sigma_t$ and $\rho_{ts} = \sigma_{ts}/(\sigma_s\sigma_t)$ where σ_{ts} denotes EU_tU_s. Denote the $T \times L$ matrix of first derivatives of M as M_θ ,

$$M_\theta(Z,\theta) = [m_{\theta 1}(Z_1,\theta)', ..., m_{\theta T}(Z_T,\theta)']', \tag{10}$$

then we have for the probit model

$$\begin{aligned} E[m_{\theta t}(Z_t,\theta)|X = x_i] &= [-\phi(x_{ti}\theta_1\theta_{2t})x_{ti}\theta_{2t}, 0, ..., \\ &\qquad -\phi(x_{ti}\theta_1\theta_{2t})x_{ti}\theta_1, ..., 0], \quad t > 0. \end{aligned} \tag{11}$$

Let ω_{tsi} be a typical element of $\Omega(x_i)$. For notational convenience let $\Phi_{ti} := \Phi(x_{ti}\theta_1\theta_{2t})$ and $\Phi_{tsi}^{(2)} := \Phi^{(2)}(x_{ti}\theta_1\theta_{2t}, x_{si}\theta_1\theta_{2s}, \rho_{ts})$. $\Phi^{(2)}(\bullet)$ denotes the cumulative distribution function of the bivariate standardised normal distribution. Hence, we obtain:

$$\omega_{tsi} = [E(Y_t - \Phi_{ti})(Y_s - \Phi_{si})|X = x_{ti}] = \begin{cases} \Phi_{ti}(1 - \Phi_{ti}) & \text{if } t = s \\ \Phi^{(2)}_{tsi} - \Phi_{ti}\Phi_{si} & \text{if } t \neq s \end{cases} \quad (12)$$

Note that $\omega_{tsi}(x_i)$ has the same sign as ρ_{ts} and that $\omega_{tsi} = 0$ if $\rho_{ts} = 0$, so that the GMM-estimator collapses to the (pooled) ML-estimator for uncorrelated errors, but generally the optimal GMM-estimator is less efficient than the ML-estimator. However, the estimation of the optimal GMM-estimator is still difficult, because it depends on the unknown correlation coefficients of Σ. Although unknown coefficients could be substituted by consistent estimates without affecting the asymptotic distribution of the estimates, obtaining them would require $(T-1)T/2$ bivariate probits, which can be cumbersome for large T. An alternative, suggested by Newey (1993), is to use nonparametric methods, such as nearest neighbour or series estimation to obtain consistent estimates of $\Omega(x_i)$.

4 Nonparametric Estimation of $\Omega(x_i)$

In the following we focus on the k-nearest neighbor (k-NN) approach, because of its simplicity. To get an easy intuition about k-NN estimation, consider the following example: Suppose there is only a finite number J of configurations for X, each containing a large number of observations N_j, then in this special case averaging the squared residuals within each subpopulation having the same $x_i (= x^q)$ e.g. $\frac{1}{N_q}\sum_{i=1}^{N_j}[y_i^q - \Phi(x^q, \tilde{\theta}_N)][y_i^q - \Phi(x^q, \tilde{\theta}_N)]'$ gives a consistent estimate for $\Omega(x_i = x^q)$ as N_q tends to infinity, where $\tilde{\theta}_N$ is a consistent, but potentially inefficient estimate. However, in most samples there are only a few observations having the same values for all explanatory variables. In this case k-NN averages locally over those residuals belonging to the k nearest (according to their similarity w.r.t. x_i) neighbors. Under regularity conditions (Newey, 1993), this will give consistent estimates of $\Omega(x_i)$ evaluated at $\tilde{\theta}_N$ and denoted by $\tilde{\Omega}(x_i)$ for each individual without the need for estimating ρ_{ts}. Thus, an element of $\Omega(x_i)$ is estimated by:

$$\tilde{\omega}_{tsi}(x_i) = \sum_{j=1}^{N} W_{tsij} m(z_{tj}, \tilde{\theta}_N) m(z_{sj}, \tilde{\theta}_N). \quad (13)$$

A weight function assigns positive weights (W_{tsij}) to those observations belonging to the k nearest neighbors ($k \leq N$), but zero weights to all other observations and the observation i itself. The weights sum to unity (Newey, 1993, eq. 4.1). Since the elements of the diagonal, ($\omega_{tsi}, t = s$) only depend on the individual indices of one time-period, we are faced to a one-dimensional estimation problem. For the off-diagonal elements, ω_{tsi}, $t \neq s$, the indices of two time-periods are important, $(x_{ti}\beta, x_{si}\beta)$, hence, the distance used to define the neighbors should refer to these two. In our Monte Carlo study we consider another possibility: The distance refers to the indices of all periods $(x_{1i}\beta, ..., x_{Ti}\beta)$ for the estimation of diagonal as well as off-diagonal elements. The advantage of this procedure is a computational one. It is much faster, because the observations have to be sorted only once for the estimation of all elements of $\tilde{\Omega}(x_i)$. Furthermore, it appears to have superior small sample properties. More details on the distance measures used and on specific weight functions employed are given in the following section.

The crucial task in k-NN estimation as in all nonparametric estimation techniques is the optimal choice of the smoothing parameter, here k. Minimising a measure of goodness-of-fit like the mean squared error (MSE) gives rise to a trade-off between squared bias and variance with respect to k. The k which minimises the *approximated* mean squared error, is, in an asymptotic sense, proportional to $n^{\frac{4}{4+d}}$ where d is the dimension of the regression, and it depends on terms, some of which are unknown in practice (cf. Mack, 1981; Bhattacharya and Mack, 1987).

In our Monte Carlo study we follow Newey (1993) in applying cross-validation for a data-driven choice of k. As Newey (p. 432) points out, the main interest is in efficient estimation of θ. Hence, he suggests a cross-validation criterion which aims at minimising remainder terms in the asymptotic theory that arise from estimation of the optimal instruments. He shows that cross-validation can be based on the difference between estimated and true moment functions:

$$\sum_{i=1}^{N} \{\tilde{A}^*(x_i) - A^*(x_i)\} M(Z, \theta^0)/\sqrt{N}. \tag{14}$$

Suppose that $\tilde{A}^*(x_i)$ and $\tilde{D}(x_i)$ denote consistent estimates of $A^*(x_i)$ and $D(x_i)$ evaluated at $\tilde{\theta}_N$. Then the resulting cross-validation function which has to be minimised is as follows (Newey, 1993, p. 433):

$$\hat{CV}(k) = \operatorname{tr}\left[Q \sum_{i=1}^{N} \tilde{R}(x_i)\tilde{\Omega}(x_i)\tilde{R}(x_i)'\right] \tag{15}$$

$$\tilde{R}(x_i) = \left\{ \frac{\partial M(z_i, \tilde{\theta}_N)}{\partial \theta} - \tilde{D}(x_i) \right\} \tilde{\Omega}(x_i)^{-1}$$
$$+ \left\{ \tilde{A}^*(x_i)[M(z_i, \tilde{\theta}_N)M(z_i, \tilde{\theta}_N)' - \tilde{\Omega}(x_i)] \right\} \tilde{\Omega}(x_i)^{-1} \quad (16)$$

Here the term $\frac{\partial M(z_i,\tilde{\theta}_N)}{\partial \theta} - \tilde{D}(x_i)$ is equal to zero, because $\frac{\partial M(z_i,\tilde{\theta}_N)}{\partial \theta}$ is identical to $D(x_i) = E\left[\frac{\partial M}{\partial \theta}(\bullet)|X = x_i\right]$, i.e. the first derivatives of $M(z_i, \theta)$ w.r.t. θ can be consistently estimated without the need for nonparametric regression. Q is in general a positive definite matrix, we always choose $Q = \sum_{i=1}^{N} D(x_i)D(x_i)'$. According to Newey's theorem 1 (p. 435) the asymptotically normal distribution and the asymptotic efficiency of $\hat{\theta}_N$ depend on several regularity conditions, for instance $k = k(N)$ has to be chosen such that $\frac{k(N)}{N} \to 0$ and $\frac{k(N)^2}{N} \to \infty$ as $N \to \infty$.

Another possibility for nonparametric estimation of $\tilde{\Omega}(x_i)$ is to use kernel regression (Carroll, 1982; Härdle, 1990, for example). However, two drawbacks of kernel regression in our context are the increased complexity of the estimation and the problem of random denominators which might produce erratic behavior (see Robinson, 1987). This problem could be solved by trimming, but this leads to a loss of efficiency by reducing the number of observations.

5 Monte Carlo Study

The aim of the following Monte Carlo study is to compare the small and large sample performance of several parametric estimators suggested in the literature with the nonparametric one discussed previously. With respect to the latter various different forms of them are plausible and the Monte Carlo study sheds some light on their different features. We use various data generating processes (DGP) for the regressors and the error process as well as different sample sizes for both panel dimensions. Before discussing the DGP and the results, the following subsection presents the estimators used in the study.

5.1 Estimators

The pooled maximum likelihood estimator which assumes independent errors over time is an obvious candidate for estimation, because of its simplicity. The estimate of θ is given by[2]

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{ti} \ln \Phi(x_{ti}\beta) + (1 - y_{ti}) \ln[1 - \Phi(x_{ti}\beta)] \qquad \text{(pooled)}$$

When considering the score function it can be seen that this is a GMM estimator combining the conditional moment restrictions given by (3) with instruments which are made of $CD(\bullet)\bar{\Omega}(\bullet)^{-1}$, where $\bar{\Omega}(\bullet)$ is identical to $\Omega(\bullet)$ on the main diagonal, but zero elsewhere. Therefore, the pooled estimator is consistent and asymptotically normal with covariance matrix given by (7). However, note that standard errors produced by a standard cross-section probit package are inconsistent whenever there are correlations of the error terms over time. Another popular method for the estimation of the panel probit model is the maximum likelihood estimator under the assumption that the error terms have a pure random effects structure, i.e. they are equicorrelated. The likelihood function can be written as follows:

$$\begin{aligned} L(y, x, \theta) &= \frac{1}{N} \sum_{i=1}^{N} \ln \int_{-\infty}^{+\infty} \sum_{t=1}^{T} \{\Phi(x_{ti}\beta + \delta c)^{y_{ti}} \\ &\quad + [1 - \Phi(x_{ti}\beta + \delta c)]^{(1-y_{ti})}\} \phi(c) dc, \end{aligned}$$

c and δ denote the random effect normalised to be standard normally distributed and its positive coefficient. Butler and Moffitt (1982) suggest an efficient method to evaluate the integral numerically, such that the estimator is given by:

$$\begin{aligned} \hat{\theta}_N &= \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ln \sum_{t=1}^{T} \sum_{v=1}^{V} \{\Phi(x_{ti}\beta + \delta c_v)^{y_{ti}} \\ &\quad + [1 - \Phi(x_{ti}\beta + \delta c_v)]^{(1-y_{ti})}\} w_v \qquad \text{(ML-RE)} \end{aligned}$$

c_v and w_v are optimally chosen evaluation points and weights. In the Monte Carlo study V is set to 5 as a compromise between computational speed and numerical accuracy (see Guilkey and Murphy, 1993, for more Monte Carlo results). When the assumed error structure is the true one this estimator is consistent and

[2] For the sake of notational convenience we omit the case of heteroscedasticity over time and set all error variances to unity.

asymptotically efficient. In the Monte Carlo study the standard errors are computed in the 'robust' way suggested by White (1982). There are no theoretical results with respect to the robustness of this estimator for true error structures which do not produce equicorrelations.

The third estimator considered is the *sequential* estimator suggested by Chamberlain (1980, 1984). The idea is as follows: In a first step a single probit is estimated for each cross-section. After computing the joint covariance matrix of all first step estimates, a minimum distance procedure is used to impose the coefficient restrictions due to the panel structure to obtain more efficient estimates. Breitung and Lechner (1994) show that this estimator is asymptotically more efficient than the *pooled* estimator, but their Monte Carlo results also indicate that there may be serious small sample problems due to the large number of first step estimates and their associated large covariance matrix.

Now we turn to the GMM estimators explicitly based on the conditional moment restrictions given in (3). All GMM estimators to be discussed in the following (and *pooled* and *sequential*) are consistent independently of the true covariance matrix of the error terms, but differ with respect to asymptotic efficiency.

The first estimator (*infeasible*) is only used as a benchmark for the others. It is computed using (3) and the true optimal instruments $A^{*}(x_i)$ evaluated at the true coefficients which are known in a Monte Carlo study. Breitung and Lechner (1994) consider several GMM estimators which are based on different parametric instruments. The estimator which is based on the instruments $A^{*}(x_i)$ computed under the assumption of random effects with a small variance (*small sigma*) of the random effect relative to the total error variance (*GMM-SS*), dominated the other GMM estimators (for details see Breitung and Lechner, 1994). Hence, we report only results for this 'best' estimator to see whether it can be improved by the use of nonparametric methods. *GMM-SS* and all nonparametric estimators depend on a preliminary consistent estimate $\tilde{\theta}_N$. In the Monte Carlo study this is always the pooled probit estimate.

Several different variants of nonparametric estimators are considered. Instead of using the conditional moments given in (3) and denoted by *NP*, one could also use the following scaled moments:

$$m_t^W(Z_{ti},\theta) = \frac{Y_{ti}-\Phi_{ti}}{\sqrt{\Phi_{ti}(1-\Phi_{ti})}} \qquad \text{(WNP)}$$

The conditional variance of the moments given by (3) is heteroscedastic across individuals, because it depends on explanatory variables, whereas the version given by (*WNP*) leads to the following conditional covariance of the moments:

$$\omega_{tsi} = [E(m_t^W m_t^{W\prime})|X = x_{ti}] = \begin{cases} 1 & \text{if } t = s, \\ \frac{\Phi_{tsi}^{(2)} - \Phi_{ti}\Phi_{si}}{\sqrt{\Phi_{ti}(1-\Phi_{ti})\Phi_{si}(1-\Phi_{si})}} & \text{if } t \neq s. \end{cases} \tag{17}$$

The latter is homoscedastic on the main diagonal and this could lead to small sample improvements. The matrix $D(\bullet)$ has to be changed accordingly. We also consider versions of *NP* and *WNP* for which only the off-diagonal elements of $\Omega(x_i)$ are estimated nonparametrically, and the main diagonal elements are set either to unity (*WNP*) or to $\Phi_{ti}(1 - \Phi_{ti})$ (*NP*). However, this leads to numerical problems in some cases of *NP*, because some eigenvalues of $\tilde{\Omega}(x_i)$ are very small or even negative.

The *WNP* and *NP* estimators can either be constructed according to (13) which estimates every single element of $\Omega(x_i)$ (*indiv*) or alternatively by estimating all elements of $\Omega(x_i)$ jointly (*joint*). The latter is motivated by computation time, the larger T the faster *joint* relative to *indiv*. The *joint* estimator is given by:

$$\tilde{\omega}_i(x_i) = \sum_{j=1}^{N} W_{ij} M(z_j, \tilde{\theta}_N) M(z_j, \tilde{\theta}_N)'. \tag{18}$$

In case that the estimate of $\Omega(x_i)$ is not positive definite, the smoothing parameter k is increased according to the grid of ks given for cross-validation until a positive definite estimate of $\Omega(x_i)$ is obtained. If the use of the largest possible k still does not produce a positive definite estimate, the values of the main diagonal are doubled until the matrix becomes positive definite. The latter way of proceeding is also applied for the versions of *NP* and *WNP* which only estimate the off-diagonal elements of $\Omega(x_i)$.

Three weight functions suggested by Stone (1977, p. 600) which fulfill the necessary regularity conditions are used. Suppose that the observations are ordered according to their distance to observation 'i'. The uniform weight function (*uniform*) which is the simplest one is then given by:

$$W_{tsij}(W_{ij}) = \begin{cases} 1/k & 1 \leq j \leq k, \\ 0 & j > k. \end{cases}$$

Smoother versions are the following quadratic (*quadr*) and triangular weight functions:

$$W_{tsij}(W_{ij}) = \begin{cases} (k^2-(j-1)^2)/[k(k+1)(4k-1)/6] & 1 \le j \le k, \\ 0 & j > k. \end{cases}$$

$$W_{tsij}(W_{ij}) = \begin{cases} (k-j+1)/(k(k+1)/2) & 1 \le j \le k, \\ 0 & j > k. \end{cases}$$

The Monte Carlo results for these smoother versions are nearly indistinguishable, so the results for the triangular weights are omitted from the tables. Finally, it remains to define the distance between two observations. First note that the distance should only depend on the linear indices $\Psi_{tsij} = [(x_{ti} - x_{tj})\beta, (x_{si} - x_{sj})\beta]$ for *indiv* and $\Psi_{ij} = [(x_{1i} - x_{1j})\beta, ..., (x_{Ti} - x_{Tj})\beta]$ for *joint*. The distance is defined by $\xi_{tsij} = \Psi_{tsij} W_{ts}^{x\beta} \Psi'_{tsij}$ or $\xi_{ij} = \Psi_{ij} W^{x\beta} \Psi'_{ij}$. If not indicated otherwise, $W^{x\beta}$ and $W_{ts}^{x\beta}$, respectively, are set to unity. However, we also use the inverse of the covariance matrix of the linear indices (S_x^{-1}) or their variance. Since there does not appear to be any significant difference between the two, the latter is omitted from the tables. The grid of ks for the cross-validation is fixed to $k = N^\eta, \eta = 0.67, 0.72, 0.77, 0.82, 0.87, 0.92, 0.97$.

5.2 Data Generating Processes and Performance Measures

The data generating processes (DGP) considered crudely mimic situations common in applied microeconometric work, e.g. regressors and error terms are both correlated over time, and dummy explanatory variables are part of the regressors. All DGP's can be summarised in the following equations:

$$\begin{aligned}
& y_{ti} = 1(\beta^C + \beta^D x_{ti}^D + \beta^N x_{ti}^N + u_{ti} > 0), && \\
& x_{ti}^D = 1(\tilde{x}_{ti}^D > 0), && P(\tilde{x}_{ti}^D > 0) = 0.5, \\
& x_{ti}^N = \gamma^x x_{t-1,i}^N + \gamma^t t + \tilde{x}_{ti}^U, && \tilde{x}_{ti}^U \sim \text{uniform}(-1,1), \\
& u_{ti} = \delta c_i + \varepsilon_{ti}, && c_i \sim N(0,1), \\
& \varepsilon_{ti} = \alpha \varepsilon_{t-1,i} + \sigma \tilde{\varepsilon}_{ti}, && \tilde{\epsilon}_{ti} \sim N(0,1), \\
\text{or} \quad & u_{ti} = 0.5(\tilde{\varepsilon}_{ti} + \tilde{\varepsilon}_{t-1,i}), && \\
& i = 1, ..., N, \quad t = 1, ..., T. &&
\end{aligned} \tag{19}$$

$(\beta^C, \beta^D, \beta^N, \gamma^x, \gamma^t, \delta, \alpha, \sigma)$ are fixed coefficients. All random numbers[3] are drawn independently over time and individuals. The first regressor is an indicator variable which is uncorrelated over time, whereas the second regressor is a smooth variable with bounded support. The dependence on lagged values and on a time trend induces a correlation over time. This type of regressor is suggested by Nerlove (1971). The error terms may exhibit correlations over time due to an individual specific effect as well as a first order autoregression or a moving average process. In order to diminish the impact of initial conditions, the dynamic processes start at $t = -10$ with $x^N_{t-11,i} = \varepsilon_{t-11,i} = 0$. T is set to 5 and 10, and N to 100, 400 and 1600 in order to study the behaviour in fairly small and really large samples. Since all estimators are $\sqrt{N}$-convergent, the standard errors for the larger sample size should be half of the next smaller sample.

Table 1 and Table 2 contain some statistics for the various DGP's used in the estimations. All DGP's have the common feature that the unconditional mean of the indicator variable is close to 0.5 in order to obtain maximum variance and thus to contain maximum information about the underlying latent variable. For ease of notation let $\mu_{ti} = \beta^C + \beta^D x^D_{ti} + \beta^N x^N_{ti}$. Table 1 gives some summary statistics for the part of the DGP related to the regressors. The coefficients γ^x and γ^t are used to generate different correlation patterns of μ_{ti} over time. Most of the simulations are based on the second configuration, the first and last one are merely considered as extreme cases.

Table 2 contains similar statistics for the error terms. The first one is the case of uncorrelated errors, the third one adds a classical individual specific random effect, and the second and fourth ones generalise the equicorrelation pattern by adding a first order autoregressive process. The fifth one is a moving average process where correlation patterns die out after one period, and finally there is an AR(1) with a negative coefficient so that the signs of the correlations alternate. Depending on T, 500 or 1000 replications (R) have been performed.

Table 3 contains the measures for the accuracy of the estimates used in the simulations. $\hat{\beta}_r$ denotes the estimate of the true value β^0 in replication 'r', and $asstd(\hat{\beta}_r)$ denotes the estimated asymptotic standard error in replication 'r'. Since in binary choice models identification is only up to scale and location, the ratio of estimated coefficients is also of interest. For the sake of brevity, the statistics related

[3] We used the random number generators RNDN and RNDU implemented in GAUSS 3.1.

to the constant term are omitted.

Since there may be concerns that the expectation and the variance of the estimates may not exist in finite samples for all estimators used, Table A.1 in appendix A presents the upper and lower bound as well as the width of the central 95% quantile of estimates based on the Monte Carlo simulations and on the asymptotic normal approximation using the average of the estimated asymptotic covariance matrices of the coefficients.

5.3 Results

For the DGPs with independent errors or an AR(1) process combined with random effects (Tables 4, 5 and A.1), several GMM-estimators with nonparametric estimation of the covariance matrix are computed for $N = 100$ and $N = 100, 400$, respectively. The results do not reveal much difference within the groups of estimators *NP* and *WNP*, however, there are substantial differences between them. (Although not reported in the tables, all estimates are almost unbiased.)

The estimators with scaled moments (*WNP*) have lower root mean squared errors (RMSE), where the differences are stronger w.r.t. the single coefficients than w.r.t. the coefficient ratio. For $N = 100$, *WNP* is sometimes involved with a somewhat larger bias of asymptotic standard errors, with *WNP-indiv-quadr* as the extreme case. Increasing the sample size N, the performance of the estimators is generally improving, becoming closer to that of the asymptotically optimal GMM-estimates given by the benchmark of the *infeasible GMM-IV*. For independent errors (Table 4), some estimates even go beyond this benchmark, a result which is due to some outliers in case of the *infeasible GMM-IV* as the inspection of the quantiles and of the minima and maxima of the estimates reveals. For $N = 100$ and $T = 5$, for instance, there is a large discrepancy between the 5%-quantile and the minimum estimate of the coefficients. Increasing the time-dimension T generally improves the results w.r.t. RMSE and bias of asymptotic standard errors.

For each replication of the Monte Carlo the smoothing parameter k is chosen via cross-validation. The distribution of the choice of k shows a tendency to choose a relatively large k. This tendency is slightly stronger for the joint than for the individual measure of distance. The same is true when increasing the time periods from 5 to 10. Both cases represent an increase in the dimension of the nonparametric estimation. *WNP* has even a stronger tendency to choose a large k than *NP*. An intuitive explanation is that the diagonal elements of the covariance for *WNP* have no bias (they are all equal to one). With increasing k the variance of the estimation is reduced. If these elements have strong weights in the cross-validation, the effect of decreasing the variance by large k might dominate the bias-variance trade-off important for the off-diagonal elements, and hence, a large (or the largest) k minimises the cross-validation function. However, replacing alternatively the diagonal elements of the covariance matrix by their parametrically estimated values and using the off-diagonal elements for the cross-validation only (*WNP-joint-uniform-no.d.*) does not seem to bring efficiency gains in contrast to the other *WNP*-methods.

The choice of the weight function for the k-NN approach, uniform or quadratic, turns out to be of minor importance. Furthermore, it does not seem to matter if the distance is measured individually or jointly over the indices $x_{ti}\beta$, although the results of the latter are more stable in the small sample. Comparing the results of Table 5 and Table A.1 we find that there are no substantial qualitative differences. Furthermore, the confidence bounds are symmetric around the true value, which implies that, at least for its tail, there is no concern about a severely asymmetric distribution.

Given these results, we choose within the class of *WNP* estimators the one with uniform weights and joint distance measure *WNP-joint-uniform* as the preferred one, because it has the simplest form of weights, it seems to be more robust w.r.t. the asymptotic standard errors than for instance the *WNP-indiv-quadr*, and saves calculation time compared to an estimator with individual distance measure.

Turning now to the parametric estimators, the following results are obtained: For the case of independent errors (Table 4) *pooled probit* is the maximum likelihood estimator. However, the *small sigma* method is already in small samples as good as *pooled probit*.

For the DGPs which have several forms of correlation of the error terms, *ML-RE* and *GMM-SS* show best results w.r.t. the RMSE. In some cases *ML-RE* is involved with a relatively high downward bias of the asymptotic standard errors for $N = 100$ (see for instance Table 7). One drawback is that it may produce inconsistent estimates when the covariance of the error terms does not follow a random effects structure as supposed in Table 6. The calculation of *ML-RE* is very computertime-consuming compared to the other methods, and often convergence cannot be reached, especially for $T = 10$.

Chamberlain's *sequential* estimator produces quite large RMSE compared to the other parametric methods for the sample size of $N = 100$, but improves with increasing N, performing however still worse than *ML-RE* and *GMM-SS*. A striking feature is the large bias of the asymptotic standard errors for $N = 100$, which is even larger for $T = 10$ than for $T = 5$. Thus, *sequential* needs quite a large sample size N in order to obtain good results. Note that the asymptotic efficiency gains for this estimator compared to *pooled*, for example, depend on the accurate estimation of the covariance matrix in the first step estimates, which then has to be inverted. In our Monte Carlo example this matrix has the dimension 15 for $T = 5$ and 30 for $T = 10$. However, the asymptotic standard errors ignore the fact that this inverse may exhibit considerable variability in small samples.

Since the *GMM-SS*-results are almost always as good as those of *ML-RE* for the different DGPs and since *GMM-SS* has in addition the advantages of producing consistent estimates for all DGPs and being computationally faster, we prefer it among the parametric estimation methods considered in this context.

It is interesting to compare *ML-RE* and *infeasible GMM-IV* for the DGP with pure random effects (Table 6). While the latter estimator represents the asymptotic efficiency bound for GMM estimators based on first conditional moments only, the former is the maximum likelihood estimator for this DGP, and hence represents the efficiency bound for all consistent estimators. Comparing these two bounds, we find that they almost coincide, so that the efficiency loss of using only first order conditional moments seems to be neglegible, once the optimal instruments are found.

Comparing now *GMM-SS* and *WNP-joint-uniform*, one observes that for the DGPs presented in Tables 4 to 7 *GMM-SS* is better than *WNP-joint-uniform* for $N =$ 100. For larger N the differences in RMSE and bias of asymptotic standard errors are very small and the ranking is inconclusive. However, when considering covariance structures which are very different from a random effects structure implying equicorrelated errors, as presented by the MA(1) process or the AR(1) process with alternating sign of the correlation over time (Tables 8 and 9), *WNP-joint-uniform* is clearly superior to *GMM-SS*.

Thus, since in real-world applications the true DGP is unknown, two estimators can be recommended to be applied: *GMM-SS* as the method which performs best among the parametric estimators and *WNP-joint-uniform* with nonparametric estimation of the covariance matrix. The second one might even be preferred since it shows almost as good (or better) results as *GMM-SS* for the DGPs presented in Tables 4 to 7, but seems to be more appropriate to deal with a wider class of structures of the covariance matrix.

[insert Tables 1 to 9 about here]

6 Applications

The panel probit model is an important model in applied microeconometrics. For example, it is the base of the econometric analysis in the paper by Laisney, Lechner and Pohlmeier (1992) which serves as one of two empirical examples to illustrate our econometric discussions. It analyses the innovative activity of firms in the West German manufacturing industry. The equation determining a firm's innovations is derived from a structural microeconomic model combining assumptions about the behavior of the individual firm with assumptions about the market structure. The dependent variable indicates whether a firm innovates or not. It is explained by firm specific variables, such as firm size relative to the market size, demand expectations of the firm, sector of the firm, and market and industry related variables such as the import share and the value added of the industry, and finally, the market share of the six largest firms in the industry as a measure of the market concentration. The sample of 1325 firms is selected from the Ifo business survey 'Konjunkturtest'. The firms are observed every year over the period from 1984 to 1988. The second study is by Pfeiffer and Pohlmeier (1992). The authors investigate the individual determinants of self-employment within a structural model of discrete choice under uncertainty. The choice equations are derived from the assumption that agents maximise a Von Neuman-Morgenstern utility function. The observed dependent variable indicates whether an individual is self-employed in period t or not. This is explained by economic variables measuring expected relative risks (unemployment and bankruptcy) and expected relative incomes in self-employment vs. dependent work. Additionally, individual heterogeneity may influence this choice. The observed components of the heterogeneity include human capital type variables, social and family background variables, and variables taking account of regional and institutional aspects. The study is based on a sample of 1926 working men selected from the German Socio-Economic Panel (GSOEP). All individuals are observed yearly from 1984 to 1989.

Table 10 contains the results of the estimations referring to the first example, which is called in the following innovation probit or INP. The results of the second example, the self-employment probit or SEP, are given in Table 11. Some descriptive statistics for both studies are shown in appendix B. For more details on the models used and the results of other specifications the reader is referred to the original papers. For the sake of simplicity and comparability with the Monte Carlo results we present only results for the model with free time effects but for which the error terms are restricted to have the same variances in each period (normalised to unity). The estimators applied to both examples include the one used in the original papers (*sequential*), the simplest one (*pooled*), the Maximum Likelihood estimator under the assumption of random effects with five evaluation points as used in the Monte Carlo study (*ML-RE 5*) and with 20 evaluation points (*ML-RE 20*), the best parametric GMM estimator (*GMM-SS*) and finally, the simplest (*WNP-joint-uniform*) and the computationally most involved (*WNP-indiv-quadratic-*S_x) of the weighted nonparametric estimators. Furthermore, we compute two versions of the asymptotic t-values for the pooled estimator: The first ones (denoted by *t-val*) ignore the possible correlations over time, and are the same that would be obtained by using a standard software package for cross-section probit estimation. The second ones are computed using the correct GMM-formula. They are comparable to the covariance matrices used in the Monte Carlo study.

Let us start by comparing the results implied by the two different ways to compute the covariance matrices of *pooled*: Not surprisingly, the t-values ignoring the intertemporal correlations are generally (with two exceptions for SEP) larger than the t-values taking account of these correlations. Therefore, for SEP, several variables which are insignificant appear to be significant when using the 'wrong' t-values.

According to Table 10 the results of the various estimators are very similar and lead to the same conclusions for INP. As expected from the Monte Carlo results, the t-values for *sequential* are somewhat higher than the other t-values and probably biased upward. Increasing the number of integration points of *ML-RE* leads to estimates which are more similar to those of the other approaches. The estimated value of δ is about 1 (std.err.: 0.04) in both cases, which implies a correlation of the errors over time of about 0.5. Figure B.1 gives the shape of the cross-validation function for three different WNP estimators (*WNP-joint-uniform-no d.* is omitted from the tables, because it almost coincides with *WNP-joint-uniform*). Although there appear to be several local minima for the *WNP-joint*-estimators, there is a clear minimum at $k = 400$ for both cases. Note that the estimator which does not estimate the main diagonal of $\Omega(x_i)$ lies above the one which does estimate the diagonal and is more erratic than that one, because of the correction made (increasing the main-diagonal) when the estimate of $\Omega(x_i)$ is not positive definite. The minimum value of the cross-validation function for *WNP-indiv-uniform-no d.* is at $k = 1200$.

Comparing the efficiency of the estimators by their estimated asymptotic standard errors (ignoring *sequential*, because of the potential downward bias), it appears that the *WNP*s dominate the other estimators for most coefficients. However, it seems that for the INP there is only a small efficiency loss when using *pooled*, which is the simplest of all estimators.

The type of results for SEP differs very much from those for INP. Using different estimators results in different conclusions leading to the overall impression that none of these results can really be trusted. The effect of increasing the number of integration points of *ML-RE* is much more pronounced than in Table 10. The estimated value of δ is about 3.6 (std.err.: 0.3, implied correlation: 0.93) for *ML-RE 5* and about 6.36 (std.err.: 0.4, implied correlation: 0.98!) for *ML-RE 20*, which hints at a very large correlation of the errors over time. Furthermore, three variables become insignificant, and two variables which are completely insignificant for *ML-RE 5*, are highly significant for *ML-RE 20*. This makes us wonder what might happen if the number of evaluation points is increased even further. Two WNP methods lead to similar results. The minimum of the cross validation function is $k = 1375$ for *WNP-joint-uniform* and $k = 1920$ for the *WNP-indiv-quadratic-Sx*. The third WNP estimator does not converge properly. An efficiency comparison of the different estimators does not appear to be sensible, because the coefficient values are too different across different estimators.

The results of Table 11 are very puzzling and deserve some explanation. The basic question is whether we observe only a *small sample* phenomenon, or whether the estimators have different probability limits, or whether both problems appear at the same time. Although the number of observations is rather large, it could be a small sample problem for the following reasons: The mean of the dummy variable to explain (e.g. the number of self-employed) is only 9% and the number of regressors is rather large and contains some dummy variables which also have a small mean. This leads to small cells for the cross-product of some explanatory variables with the dependent variable, which could lead to a deterioration of the performance and to an instability of the estimators. Furthermore, if the estimate of the correlation of the error terms of *ML-RE* can be taken seriously, then the correlation is almost unity. However, such a high correlation taken together with the fact that the index $x\beta$ varies only very little over time, because almost all important variables are time constant, leads to severe problems for the estimation of the inverse of $\Omega(x_i)$. This could explain the performance of the GMM-estimators which all use this matrix as instrument in one form or another. But there may be as well concerns that the estimators may have a different probability limit: Lechner (1995) proposes specification tests for the panel probit model and uses the same empirical example. He finds that the specification is questionable and that it appears very likely that there are endogeneity problems for at least one explanatory variable. The facts that the *GMM* and *ML-RE* estimators rely on the assumption of strict exogeneity and *sequential* and *pooled* still rely on weak exogeneity of the regressors, and that the different groups of estimators use the potentially invalid instruments in a different way, could be sufficient for them to converge to different limits (which suggests to construct a Hausman-type test to test strict exogeneity). However, all this remains speculation. Although taking into account that the *pooled* estimator does not need estimates of complicated weighting matrices and is more robust than *GMM* and *ML-RE* with respect to strict exogeneity and therefore seems preferable to the other estimators for this study, the main conclusion seems to be that the differences of the estimators indicate problems with the specification itself. This shows another useful purpose of using several similar estimators for the same empirical study.

[insert Tables 10 and 11 about here]

7 Conclusions

Applying maximum likelihood estimation to non-linear models on panel data requires high-dimensional numerical integration over probability density functions due to intertemporal correlations of the error terms. One alternative analysed in this work is to combine the generalised method of moments (GMM) with the nonparametric estimation of the instrument matrix as proposed by Newey (1993). The result is an easily computable estimator with good small sample properties and only small efficiency loss compared to full information maximum likelihood.

Although for some data generating processes and especially in small samples, maximum likelihood with numerical integration as suggested by Butler and Moffitt (1982) performs better than the other estimators considered in the Monte Carlo study, it is often involved with convergence problems when sample sizes N or T are increasing and when data generating processes of the error terms are getting more complicated. In the latter case, for instance assuming an MA(1) process or an AR(1) process with alternating signs of the correlation coefficient, the generalised method of moments with nonparametric estimation of the covariance matrix (*WNP*) is superior to all other methods.

Another variant of GMM already analysed in Breitung and Lechner (1994), *GMM-SS*, also gives good results throughout the considered data generating processes, however, it is outperformed by *WNP* when the intertemporal correlation of the error terms is assumed to be more complicated.

Since in real-world applications the covariance matrix of the error terms is unknown one might prefer an estimator which is known to give good results with respect to efficiency independently of the underlying data generating process. Therefore, the GMM method with nonparametric estimation of the covariance matrix seems to be a very competitive candidate.

We apply the various methods to two microeconometric models. In the first case the results of the estimators are very similar, but in the second one they diverge substantially. We consider the latter as evidence for possible small sample problems and a misspecifaction of the conditional moment function.

One aim of future research should be to analyse if the same results can be obtained for other non-linear models such as tobit. Another topic of interest might be the comparison of other nonparametric methods with the k-nearest neighbor approach within this context as well as to consider other approaches for choosing the smoothing parameter.

Appendix A: Additional Monte Carlo Results

In the following the bounds and the widths of the 95% confidence intervals for one of the data generating processes (DGP=2, Table 5) are given.

[insert Table A.1 about here]

Appendix B: Additional Information about the Applications

The two tables show descriptive statistics of the variables used in the two applications. The cross-validation functions obtained for the k-nearest neighbor estimation of the covariance matrix are depicted in Figures B.1 and B.2.

[insert Tables B.1 and B.2 about here]

[insert Figures B.1 and B.2 about here]

References

Avery, R., Hansen, L. and Hotz, V. (1983). Multiperiod probit models and orthogonality condition estimation, *International Economic Review* **24**: 21–35.

Bhattacharya, P. K. and Mack, Y. P. (1987). Weak convergence of k-nn density and regression estimators with varying k and applications, *The Annals of Statistics* **15**(3): 976–994.

Breitung, J. and Lechner, M. (1994). GMM-estimation of non-linear models on panel data. Discussion paper No. 500-94, University of Mannheim.

Butler, I. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model, *Econometrica* **50**: 761–764.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models, *The Annals of Statistics* **10**(4): 1224–1233.

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

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics* **34**: 305–334.

Guilkey, D. and Murphy, J. (1993). Estimation and testing in the random effects probit model, *Journal of Econometrics* **59**: 301–317.

Hajivassiliou, V. (1993). Simulation estimation methods for limited dependent variable models, *in* G. Maddala, C. Rao and H. Vinod (eds), *Handbook of Statistics*, Vol. 11, North-Holland, Amsterdam.

Hansen, L. (1982). Large sample properties of generalized methods of moments estimators, *Econometrica* **50**: 1029–1055.

Härdle, W. (1990). *Applied Nonparametric Regression*, number 19 in *Econometrics Society Monographs*, Cambridge University Press.

Horowitz, J. (1993). Semiparametric and nonparametric estimation of quantal response models, *in* G. Maddala, C. Rao and H. Vinod (eds), *Handbook of Statistics*, Vol. 11, North-Holland, Amsterdam.

Laisney, F., Lechner, M. and Pohlmeier, W. (1992). Innovation activity and firm heterogeneity: Empirical evidence from West Germany, *Structural Change and Economic Dynamics* **2**: 301–319.

Lechner, M. (1995). Some specification tests for probit models estimated on panel data, *Journal of Business and Economic Statistics*. forthcoming.

Lechner, M. and Breitung, J. (1995). Some GMM estimation methods and specification tests for nonlinear models, *in* L. Mátyás and P. Sevestre (eds), *The Econometrics of Panel Data*, Kluwer Academic Publishers, Netherlands. 2nd ed., forthcoming.

Mack, Y. P. (1981). Local properties of k-NN regression estimates, *SIAM Journal of Algebraic and Discrete Methods* **2**(3): 311–323.

Manski, C. F. (1988). Identification of binary response models, *Journal of the American Statistical Association* **83**: 729–738.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross sections, *Econometrica* **39**: 359–383.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models, *Econometrica* **59**: 809–837.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions, *in* G. Maddala, C. Rao and H. Vinod (eds), *Handbook of Statistics*, Vol. 11, North-Holland, Amsterdam, chapter 16.

Pfeiffer, F. and Pohlmeier, W. (1992). Income, uncertainty and the probability of self-employment, *Recherches Economiques de Louvain* **58**: 265–281.

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form, *Econometrica* **55**(4): 875–891.

Stone, C. J. (1977). Consistent nonparametric regression, *The Annals of Statistics* **5**(4): 595–645.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.