TI 2015-112/V
Tinbergen Institute Discussion Paper

A New View on Panel Econometrics. Is Probit feasible after all ?

Bernard M.S. Van Praag

Faculty of Economics and Business, University of Amsterdam, and Tinbergen Institute, the Netherlands.

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

A New View on Panel Econometrics.
Is Probit feasible after all ?

by

Bernard M.S. Van Praag

Amsterdam School of Economics, Tinbergen institute, CESifo, IZA

Amsterdam, August 19, 2015

b.m.s.vanpraag@uva.nl

Abstract

A New View on Panel Econometrics. Is Probit feasible after all?

Mundlak (1978) proposed the addition of time averages to the usual panel equation in order to remove the fixed effects bias. We extend this Mundlak equation further by replacing the time-varying explanatory variables by the corresponding deviations from the averages over time, while keeping the time averages in the equation. It appears that regression on this extended equation provides simultaneously the within- and the in- between- estimator, while the pooled data estimator is a weighted average of the within and in-between estimator. In Section 3 we introduce observed and unobserved fixed effects In Section 4 we demonstrate that in this extended setup Probit - estimation on panel data sets does not pose a specific problem. The usual software will do. In Section 5 we give an empirical example**.**

Keywords: Panel data estimation techniques, ordered probit, fixed effects-estimator, within-estimator, pooled regression, between-estimator.

JEL code: C23, C25.

1.Introduction

An important field of econometrics is panel econometrics. In most textbooks separate chapters are devoted to panel analysis, and several textbooks like Baltagi (2006) and Wooldrich (2002) are almost exclusively focusing on panel econometrics. The main difficulty in panel econometrics is that the traditional linear model $y_{it} = \beta' x_{it} + \gamma' z_i + \varepsilon_{it}$ offers at least three ways of estimating the coefficients. We may apply OLS on the pooled data (y_{it}, x_{it}) , we may apply OLS on the differences $(\ddot{y}_{it}, \ddot{x}_{it})$ from the mean (within), and we may apply OLS on the averages over time $(\overline{y}_{it}, \overline{x}_{it})$ (in -between). The empirical problem is that the results of these three estimation methods for β may yield very different results. Mostly, it is explained by the existence of unobserved individual effects α_i that are correlated with the x- variables. This is actually a special case of omitted variables bias and/or correlation between the x and the errors.

Mundlak (1978) and Chamberlain (1984) show that, if the α_i are correlated with x, they may be decomposed into a linear function of the time averages $\delta' \overline{x}_i$ and an independent error. Hence, they add the time averages as a second set of explanatory variables and take for basic model $y_{it} = \beta' x_{it} + \delta' \overline{x}_i + \gamma' z_i + \varepsilon_{it}$. If the individual effects α_i happen to be uncorrelated with *x* and/or *z,* they are called random effects.

In this paper we go one step further and replace the x_{it} in the last equation by $\ddot{x}_{it} = x_{it} - \overline{x}_i$, yielding the extended model $y_{it} = \ddot{\beta}' \ddot{x}_{it} + \overline{\beta}' \overline{x}_i + \gamma' z_i + \varepsilon_{it}$. This implies a loosening of the implicit assumption of the traditional model that $\ddot{\beta} = \overline{\beta}$. It may be true or not and empirical testing will decide on it.

In this paper we will demonstrate in Section 2 that regression on this extended model will provide us simultaneously with an estimator of $\ddot{\beta}$,which is identical to the within- or fixed- effects estimator $\hat{b}_{within}$, and with an estimator of $\overline{\beta}$ which equals the in-between estimator $\hat{b}_{inbet}$. When we apply regression under the constraint that $\hat{b}_{within} = \hat{b}_{inbet} = \hat{b}$, then $\hat{b}$ will be a weighted average of the

estimators $\hat{b}_{within}$ and $\hat{b}_{inbet}$. In Section 3 we introduce the observed time-constant variables Z_i and the (possibly correlated) unobserved fixed effects α_i. In Section 4 we demonstrate that this approach works for the Probit- version of the model as well. Contrary to the idea that the Probit model is impossible to estimate for panel data and that consequently we have to take recourse to a Logit specification, we show that such estimation is possible on the extended model by standard software. In Section 5 we give an empirical example. In Section 6 we draw some conclusions.

The upshot of this is that linear panel analysis, based on the model $y_{it} = \ddot{\beta}'\ddot{x}_{it} + \overline{\beta}'\overline{x}_i + \gamma' z_i + \varepsilon_{it}$ does not pose special 'panel problems', although we may still encounter the usual problems found in classical multivariate statistics in general and in linear econometrics in particular. It is therefore that we argue that panel analysis can be re-integrated in traditional classical econometrics, when we replace the traditional model $y_{it} = \beta' x_{it} + \gamma' z_i + \varepsilon_{it}$ by the extended model $y_{it} = \ddot{\beta}'\ddot{x}_{it} + \overline{\beta}'\overline{x}_i + \gamma' z_i + \varepsilon_{it}$.

2. Comparison of the traditional estimators with the estimators of the extended model.

First, we consider the linear panel data model in its most simple form. This is

$$y_{it} = \beta_1 x_{1,it} + ... + \beta_K x_{K,it} + \beta_0 + \varepsilon_{it} \qquad (2.1)$$

where *i* (*i=1,…,N*) stands for the *i*th observation unit and *t* (*t=1,…,T*) for the different observation moments. The errors are assumed to be mutually independent. Moreover, we assume for the errors zero expectation and homo-skedasticity. The error covariance matrix is $\Sigma_\varepsilon = \sigma_\varepsilon^2 . I_{NT}$.

The *T* observations per observation unit i are denoted as $\left(y_{it}, x_{1,it},...,x_{K,it},1\right)$ or for short as $\left(y_{it}, x_{it}\right)$. We notice that x_{it} is a (K+1)-vector, where we include the constant 1.

If we consider all *T* observations for unit *i* we store them as the *T* x(*K+2*)- matrix $\left(y_i, x_{1,i},...,x_{K,i}\right)$ or $\left(y_i, X_i\right)$. All individual observations, including the constant, are stored in the *NT x* (*K+2*)- matrix $\left(y\ , X\ \right)$.

In this paper we will use semi-definite symmetric (*TxT*)- matrices Σ, M and their generalized inverses .The generalized inverse will be denoted by Σ^+ and it is defined as the matrix with the same eigenvectors as those of Σ; they share the zero eigenvalues, while their corresponding non-zero eigenvalues are the reciprocals of those of Σ. While those matrices are generally *TxT*-matrices, we will use incidentally block-diagonal (*NTxNT*) -matrices like $\mathbf{\Sigma}_\varepsilon = I_N \otimes \Sigma_\varepsilon$. They will be denoted by bold symbols.

Let us denote the averaging matrix by M , where $M = \frac{1}{T}\iota\iota'$ is a (TxT) –matrix with each cell equaling 1/T. The matrix *M* is idempotent of rank 1. We have for the vector of averages $\overline{X}_i = MX_i$. Similarly, the demeaning procedure can be described by the matrix $(I - M)$ with rank (*T-1*) and we have for the deviations from the mean $\ddot{X}_i = (I - M)X_i$.

Due to the fact that *M* is idempotent, we have $M(I - M) = O$ and hence $\ddot{X}_i'\overline{X}_i = X_i'(I - M)MX_i = O$. In a similar way it may be shown, for example, that $\ddot{X}'\ddot{y} = \ddot{X}'(\ddot{y} + \overline{y}) = \ddot{X}'\ddot{y},\ \ddot{X}'\overline{\varepsilon} = o$. In short, the deviations are not correlated with the time- constant variables. Similarly, we find that $\ddot{X}'\Sigma = \ddot{X}'(\Sigma_{\overline{\varepsilon}} + \Sigma_{\ddot{\varepsilon}}) = \ddot{X}'(M\Sigma M\ + (I - M)\Sigma(I - M)) = \ddot{X}'\Sigma_{\ddot{\varepsilon}}$.

There are in the literature three basic methods to estimate the parameter vector β . The first method is of course OLS applied on (2.1).

The second method is the so-called *in- between* estimation. We consider the average observations over time. We denote the time averages for observation unit *i* by $(\overline{y}_i, \overline{x}_{1,i}, ..., \overline{x}_{K,i})$ and the average error by $\overline{\varepsilon}_i$. Then (2.1) implies for the averages

$$\overline{y}_i = \beta_1 \overline{x}_{1,i} + ... + \beta_K \overline{x}_{K,i} + \beta_0 + \overline{\varepsilon}_i \quad (i=1,...,N) \tag{2.2}$$

where $\Sigma_{\overline{\varepsilon}} = \frac{1}{T}\sigma_\varepsilon^2 . I_N$ .So it follows that β may be estimated by OLS on (2.2) as well. This estimator is called the *in-between estimator* $\hat{b}_{ib}$. We may rewrite (2.2) as

$$My_i = MX_i\beta + M\varepsilon_i \tag{2.3}$$

The third estimator is found by taking the differences from the means. Let us denote $\ddot{y}_{it} = y_{it} - \overline{y}_i$ and let us use the same notation for the other variables.

Then we find that there holds

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{1,it} + ... + \beta_K \ddot{x}_{K,it} + \ddot{\varepsilon}_{it} \quad (i=1,...N,\ t=1,...,T) \tag{2.4}$$

or

$$(I - M)y_i = (I - M)X_i\beta + (I - M)\varepsilon_i \tag{2.5}$$

Applying OLS (without intercept) on (2.4), yields a third estimator. It is the so-called *within-estimator* of β , which we denote by $\hat{b}_{wi}$.

We see that the error vector $\ddot{\varepsilon}_i = \left(\ddot{\varepsilon}_{i,1}, ..., \ddot{\varepsilon}_{i,T}\right)$ has a non-diagonal covariance matrix

$$\Sigma_{\ddot{\varepsilon}} = \left(I - M\right)\sigma_\varepsilon^2 . \left(I - M\right) \tag{2.6}$$

This matrix $\Sigma_{\ddot{\varepsilon}}$ is singular of rank *T-1.* This follows automatically as there are only (*T-1*) independent equations for observation unit *i*, reflecting the fact that the equation for observation (*i ,T*) follows automatically from the fact that the *T*

demeaned observations add up to zero. In the following context we might, as usual, drop the last observation for each observation unit *i,* which removes the singularity. For our following analysis we prefer to use all *T* data. Then we have to replace the usual inverse by the generalized inverse $\Sigma_{\ddot{\varepsilon}}^{+}$ which has the same eigenvectors as $\Sigma_{\ddot{\varepsilon}}$, while the non-zero eigenvalues are replaced by their reciprocals. The three OLS- estimators being consistent, they are estimating the same parameter vector β, except for the within-estimator which does not estimate the intercept β_0 .

Now we add a fourth extended model

$$y_{it} = \ddot{\beta}'\ddot{x}_{it} + \overline{\beta}'\overline{x}_{it} + \varepsilon_{it} \tag{2.7}$$

or

$$y_i = (I - M)X_i\ddot{\beta} + MX_i\overline{\beta} + M\varepsilon_i + (I - M)\varepsilon_i \tag{2.8}$$

This model leaves explicitly open that $\ddot{\beta} \neq \overline{\beta}$. If there holds equality, we are back in the traditional case, since $\beta'(\ddot{x}_{it} + \overline{x}_{it}) = \beta' x_{it}$. Hence (2.1) is a special case of the more general model (2.7). The implicit traditional assumption that $\ddot{\beta} = \overline{\beta}$ may now be tested as a hypothesis. We notice that $\ddot{X}_i$ is a (*TxK*)- matrix while $\overline{X}_i$ is a (Tx(K+1))- matrix, including the intercept.

The first term in (2.7) refers to the differences and the second term to the averages. Since both terms are uncorrelated we may apply the Frisch –Waugh-Lovell theorem (Frisch, Waugh (1933), Lovell (1963)).

Since $\ddot{x}'y = \ddot{x}'\ddot{y}$ and $\ddot{x}'\varepsilon = \ddot{x}'\ddot{\varepsilon}$ we find that regression on $\ddot{x}$ only, i.e.,

$$y_{it} = \ddot{\beta}'\ddot{x}_{it} + \varepsilon_{it} \tag{2.9}$$

will yield the estimator $\hat{\ddot{b}} \equiv \hat{b}_{within} = \left(\ddot{X}'\ddot{X}\right)^{-1}\ddot{X}'y$, identical with the result when regressing on $\ddot{y}_{it}$ as in (2.4).

Similarly, dropping the differences in (2.8), OLS yields only the in-between estimator. In short $\hat{\bar{b}} \equiv \hat{b}_{inbetween}$. Regression on (2.8) will simultaneously provide us with the within – and the in-between estimator.

We get the explicit formulae:

$$\begin{aligned} \hat{b} &= \left(X'X \right)^{-1} X'y \\ \hat{b}_{wi} &= \left(\ddot{X}'\ddot{X} \right)^{-1} \ddot{X}'y \\ \hat{b}_{ib} &= \left(\bar{X}'\bar{X} \right)^{-1} \bar{X}'y \end{aligned} \qquad (2.10)$$

where $\boldsymbol{\Sigma}_{\ddot{\varepsilon}} = I_N \otimes \Sigma_{\ddot{\varepsilon}}$ is a (*TN x TN*) – block- diagonal matrix. For $\bar{X}$ we remark that the first *T* rows are identical, and so for the second *T*-tuple, and so on. We may also interpret $\bar{X}$ in the formula (2.5) as a (*N* x(K*+1*))-matrix, and $\bar{y}$ as an *N* –vector $\bar{y}' = (\bar{y}_1, ..., \bar{y}_N)$. Then the (*TN x TN*) – block- diagonal matrix $\boldsymbol{\Sigma}_{\bar{\varepsilon}}$ = $I_N \otimes \Sigma_{\bar{\varepsilon}}$ has to be replaced by the (*N x N*) - matrix $\Sigma_{\bar{\varepsilon}}$ as well.

Since the OLS- estimators are unbiased, we may write the differences of the estimators with their expectations as $\hat{b}_{with} - \ddot{\beta} = \left(\bar{X}'\bar{X} \right)^{-1} \bar{X}'\ddot{\varepsilon}$

and $\hat{b}_{ib} - \bar{\beta} = \left(\bar{X}'\bar{X} \right)^{-1} \bar{X}'\bar{\varepsilon}$. Since the errors $\ddot{\varepsilon}, \bar{\varepsilon}$ are uncorrelated, it follows that the estimators $\hat{b}_{wi}, \hat{b}_{ib}$ are uncorrelated as well. If they are (asymptotically) normal, they are mutually independent.

The covariance matrices of the estimators are:

$$\begin{aligned}\operatorname{cov}(\hat{b}_{wi}) &= \frac{\sigma_\varepsilon^2}{NT}\left(\ddot{X}'\ddot{X}\right)^{-1} \\ \operatorname{cov}(\hat{b}_{ib}) &= \frac{\sigma_\varepsilon^2}{NT}\left(\overline{X}'\overline{X}\right)^{-1}\end{aligned} \tag{2.11}$$

where $\overline{X}$ is a (*NT x*(*K+1*))-matrix. When we take $\overline{X}$ to be a (*N x*(*K+1*)) the first factor becomes $\sigma_\varepsilon^2 \big/ N$.

It is now easy to test the equality of $\hat{b}_{wi}$ and $\hat{b}_{ib}$. We notice that $\operatorname{cov}(\hat{b}_{wi} - \hat{b}_{ib}) = \operatorname{cov}(\hat{b}_{wi}) + \operatorname{cov}(\hat{b}_{ib}) = \Sigma_{wi.ib}$. Hence, we have the test criterion $(\hat{b}_{wi} - \hat{b}_{ib})'\Sigma^{-1}{}_{wi.ib}(\hat{b}_{wi} - \hat{b}_{ib}) \sim \chi_k^2$.

The *pooled* estimator $\hat{b} = \left(X'X\right)^{-1}X'y$, where we estimate under the additional constraint $\hat{b}_{wi} = \hat{b}_{ib} = \hat{b}$, or in terms of the population parameters $\ddot{\beta} = \overline{\beta} = \beta$, may now be interpreted as a weighted sum of the within- and in-between estimator. We write the sum of squared residuals using the zero-correlation of the means and the deviations as

$$\begin{aligned}S^2 &= \sum_{i,t}(y_{it} - \beta' x_{it})^2 \\ &= \sum_{i,t}(\ddot{y}_{it} - \beta'\ddot{x}_{it})^2 + T.\sum_{i}(\overline{y}_i - \beta'\overline{x}_i)^2 \\ &= \ddot{S}^2 + T.\overline{S}^2\end{aligned} \tag{2.12}$$

Differentiation with respect to β yields the normal equation

$$\left(\ddot{X}'\ddot{X} + T.\overline{X}'\overline{X}\right)\hat{b} = \ddot{X}'\ddot{y} + T.\overline{X}'\overline{y}\ .$$

We solve for β and get

$$\hat{b} = \left(\ddot{X}'\ddot{X} + T.\bar{X}'\bar{X} \right)^{-1} \left(\ddot{X}'\ddot{y} + T.\bar{X}'\bar{y} \right) =$$

$$\left(\ddot{X}'\ddot{X} + T.\bar{X}'\bar{X} \right)^{-1} (\left(\ddot{X}'\ddot{X} \right)\left(\ddot{X}'\ddot{X} \right)^{+} \ddot{X}'\ddot{y} + \left(\ddot{X}'\ddot{X} + T.\bar{X}'\bar{X} \right)^{-1} \left(\bar{X}'\bar{X} \right)\left(\bar{X}'\bar{X} \right)^{+} .T.\bar{X}'\bar{y})$$

This is a matrix-weighted average $\hat{b} = W_{wi}\hat{b}_{wi} + W_{ib}\hat{b}_{ib}$ with weights $W_{wi} = \left(\ddot{X}'\ddot{X} + T.\bar{X}'\bar{X} \right)^{-1} \left(\ddot{X}'\ddot{X} \right)$ and $W_{ib} = \left(\ddot{X}'\ddot{X} + T.\bar{X}'\bar{X} \right)^{-1} \left(\bar{X}'\bar{X} \right)$, respectively. We may write the weights even more elegantly as $W_{wi} = \left(X'X \right)^{-1} \left(\ddot{X}'\ddot{X} \right)$ and $W_{ib} = \left(X'X \right)^{-1} \left(\bar{X}'\bar{X} \right)$.

We see that the within-estimator gets the upper weight if the inter-temporal variation is large and inter-individual variation is small, while the within – estimator becomes relatively unimportant if inter-temporal variations per individual are small, but inter-individual variations are relatively large.

If the equality $\hat{b}_{wi} = \hat{b}_{ib}$ holds, we have for their joint estimator $\hat{b}$

$$\operatorname{cov}(\hat{b}) = \frac{\sigma_{\varepsilon}^{2}}{NT} \left(X'X \right)^{-1}.$$

3. ***Observed and unobserved fixed effects.***

It is frequently found that the estimators $\hat{b}_{wi}, \hat{b}_{ib}$ yield statistically different estimates of the parameter vector β, which contradicts the model equation (2.1). The traditional way to repair this contradiction is to extend the model by adding a vector of *M* observed effects *Z* and/or an unobserved individual fixed effect α_i .The equation runs

$$y_{it} = \beta' x_{it} + \gamma' Z_i + \alpha_i + \beta_0 + \varepsilon_{it} \qquad (3.1)$$

The corresponding extended version is

$$y_{it} = \ddot{\beta}'\ddot{x}_{it} + \overline{\beta}'\overline{x}_i + \gamma' Z_i + \alpha_i + \beta_0 + \varepsilon_{it} \tag{3.2}$$

We notice that (3.2) is again a sum of two uncorrelated sets of explanatory variables, viz., $\ddot{X}$ and $(\overline{X}, Z, \alpha)$. The role of $\overline{X}$ and Z are the same. The role of α is that of an omitted variable. If it correlates with $\overline{X}$ or Z, it creates an 'omitted variable'- bias for the estimators of $\overline{\beta}$ and γ . If it does not correlate, it is just a random effect, which may be added to the in-between error yielding a composite in-between error $\eta_i = \overline{\varepsilon}_i + \alpha_i$ with corresponding co-variance matrix $\Sigma_\eta = \left(\overline{\sigma}_\varepsilon^2 + \sigma_\alpha^2\right) I$.

We may regress *y* on $\ddot{X}$ only, yielding the within-estimator. The within- estimator is estimated from the normal equation

$$\left(\ddot{X}'\ddot{X}\right)\ddot{\beta} = \ddot{X}'y \tag{3.3}$$

The within –estimator is not affected by the presence of Z and α , since $\ddot{X}, \ddot{y}, \ddot{\varepsilon}$ are uncorrelated with Z and α . For the same reason the within-estimator and its covariance- matrix is not affected by the presence of random effects.

We denote the extended matrix of explanatory variables by $\overline{X}Z = (\overline{X}, Z)$.

The in - between estimator is estimated from the normal equations

$$\begin{bmatrix} \left(\overline{X}\,'\overline{X}\right) & \left(\overline{X}\,'Z\right) \\ \left(\overline{X}\,'Z\right)' & (Z'Z) \end{bmatrix} \begin{bmatrix} \overline{\beta} \\ \gamma \end{bmatrix} = \begin{bmatrix} \left(\overline{X}\right) \\ (Z) \end{bmatrix}' \left[\overline{y} - \alpha\right] \tag{3.4}$$

where we notice that the relevant covariance matrix is $\Sigma_\eta = \left(\ddot{\sigma}_\varepsilon^2 + \sigma_\alpha^2\right) I$ .The estimators of $\overline{\beta}$ and γ are not consistent if $\left(\overline{X}\right)' \alpha \neq o$ and/or $Z'\alpha \neq o$. This is nothing else than the omitted variable bias when we apply OLS on the equation $y_{it} = \overline{\beta}'\overline{x}_i + \gamma' Z_i + \alpha_i + \beta_0 + \varepsilon_{it}$.

Notice that the pooled estimator is also affected by the presence of Z and α . This is easily shown by the fact that the *pooled* estimator, as we saw in the previous section, may be interpreted as a weighted sum of the within- and in-between -estimator. If one of the two estimators is biased, it is obvious that the weighted average will be biased as well.

Alternative interpretation

An alternative explanation for an observed difference between the estimators $\hat{b}_{wi}, \hat{b}_{ib}$ is that the difference is not caused by an imperfect specification of the model equation, but that the two estimators stand for two *different* effects.

We may interpret (3.2) as a decomposition between the *structural* time invariant effects of the average $\overline{X}_i$ $\overline{X}$ and Z , and the effect of the *temporal* deviation $\ddot{X}_{it}$. An example of an economic theory where such a model seems relevant is Friedman's (1956) permanent income theory. Consumption y_{it} is there explained by permanent income $\overline{X}_i$ and the income fluctuations $\ddot{X}_{it}$ about the mean.

As said before, whether the hypothesis $\beta_{ib} = \beta_{wi}$ is justified can be statistically tested.

4. *Application to qualitative panels.*

The present findings facilitate our approach to linear panel analysis. However, the main surprising novel result of this analysis is its consequence for the Ordered Probit analysis of panel data. In the established literature it is taken for granted that panel equations cannot be estimated by Ordered Probit. See, for instance, Chamberlain (1984), or Cameron and Trivedi (2005), who state in Section 23.4.3. of their magnificent book '' *Fixed Effects estimation is possible for the panel logit model, using the conditional MLE, but not for other binary panel models such as panel probit'.* This is based on the following reasoning.

Let the latent model be the traditional version

$$y_{it} = \beta' x_{it} + \gamma' Z_i + \alpha_i + \beta_0 + \varepsilon_{it} \qquad (4.1)$$

where the error according to the Ordered Probit – model is assumed to follow a N(0,1)- distribution. We can apply OP on (4.1) .In that case the established result is that the OP-estimators are ML-estimators; consequently, the ML-estimators $\hat{b}_{OP}$ will tend to the probability limits $\hat{b}_{OLS}$ obtained by OLS on (4.1). However, if there are fixed effects, $\hat{b}_{OLS}$ will be biased as we saw.

The usual way out in the panel- OLS-model is to look at the equation (2.4), viz.,

$$\ddot{y}_{it} = \ddot{X}_{it}'\beta + \ddot{\varepsilon}_{it} \qquad (4.2)$$

However, now we observe the y's only qualitatively as belonging to one of J ordered response categories $\left\{ \mu_{j-1} < y \le \mu_j \right\}$; we do not have the qualitative analogue of $\ddot{y}_{it}$. Hence, we cannot apply OP on (4.2).

Let us now apply OP on the latent model

$$y_{it} = \ddot{\beta}'\ddot{x}_{it} + \overline{\beta}'\overline{x}_i + \gamma' Z_i + \alpha_i + \beta_0 + \varepsilon_{it} \qquad (4.3)$$

where the error is assumed to follow a N(0,1)- distribution. When we could apply OLS on the quantitative observations y_{it} following (4.3), the within – and in-between estimators would be simultaneously and consistently be estimated on (4.3) by $\hat{b}_{wi}$ and $\hat{b}_{ib}$. It is well-known that if we only know the observations as belonging to *ordered* classes, while the errors are N(0,1)- distributed, then we may apply OP and the OP-estimators, being M L- estimators, are consistent. However, we are still assuming that errors are N(0,1)- distributed. The latter assumption is unnecessarily restrictive. If the errors have an arbitrary variance $\sigma > 0$, the estimation under the false assumption that $\sigma = 1$will yield OP-estimators which are all multiplied by $1/\sigma$, but which have the same ratio to each other.

In the qualitative panel context the observations of y_{it} are belonging to known *ordered* classes. Hence, we may apply OP on (4.3) and we estimate the parameters $\overline{\beta}, \ddot{\beta}$ or β_{ib}, β_{wi} up to an unknown proportionality factor $\psi = 1/\sigma$ (>0).

The coefficients and their covariance matrix may be estimated by OP by using traditional software. The estimators are consistent. Hence $\hat{\ddot{b}}_{OP} \approx \psi.\hat{\ddot{b}}_{OLS}$ and $\hat{\overline{b}}_{OP} \approx \psi.\hat{\overline{b}}_{OLS}$. If there are fixed effects α we encounter the same problems as in the OLS-case for the in-between and the pooled case. In case of correlation the estimators $\hat{\overline{b}}_{OP}, \hat{\gamma}_{OP}$ will be biased. However, $\hat{\ddot{b}}_{OP}$ will be consistent. It is immune for fixed effects. Again, it is possible to test whether $\hat{\ddot{b}}_{OP} = \hat{\overline{b}}_{OP}$.This is relatively easy as the estimators $\hat{\ddot{b}}_{OP}, \hat{\overline{b}}_{OP}$ are uncorrelated. This follows when we

consider the joint covariance matrix. It may be written as $\left[\sum_{i=1}^{N} w_i x_i x_i'\right]^{-1}$ (see Cameron and Trivedi, p.469), where $x_i' = \left[\ddot{x}_i', \overline{x}_i'\right]$. It is easy to see that the non-diagonal blocks in this matrix are a sum of zeroes and therefore zero themselves.

The difference between the traditional model and our model is that the existing literature states that the within- estimator β_{wi} can only be derived by regressing on the *differences* $\ddot{y}_{it}$, while we showed above that regressing on the original observations y_{it} yields the same estimator. This fact sets us free from the necessity to find a counterpart for the difference $\ddot{y}_{it}$ in the Probit –context. The impossibility to define the difference $\ddot{y}_{it}$ in the Probit –context led researchers to the conviction that estimation of a qualitative linear panel model by OP was impossible. However, since it appears that also for the quantitative model using $\ddot{y}_{it}$ is not needed at all, it follows logically that this is not necessary for OP either. Analogously to (2.8) we may estimate β_{wi} from the latent model (2.9) $y_{it} = \ddot{\beta}'\ddot{x}_{it} + \varepsilon_{it}$.

5. ***Empirical example***

As an illustration of the above we use the data of Vella and Verbeek (1998)[1].First, we estimate an equation where the variable *y* to be explained is continuously observed. Then we discretize *y* into two events: positive or negative and we explain those discrete events by Probit, assuming the same now latent model

[1] These data are downloadable from the Journal of Applied Econometrics-data archive.

equation[2]. Vella and Verbeek use data from the National Longitudinal Survey (Youth Sample). The set consists of 545 observations of young full-time working males who have completed their schooling by 1980. Annual observations of the wage are available for 1980-1987. Here we use the data to estimate a standard Mincerian wage equation with some additional explanatory variables. We use the following time-constant explanatory variables: years of schooling (School), Black and Hispanic. The time -varying explanatory variables are: a proxy for labor market experience (Experience = Age - 6 - School) and its square, being married or not (Married), and union membership (Union). Controls for region of residence are added as well. We distinguish three kinds of variables: the time constant variables *Z,* the demeaned variables $\ddot{x}$ and the averages of *X, denoted by* $\overline{x}$.

OLS-results.

The results of five different estimation methods are presented in Table 1. First we look at OLS on $\Delta x = \ddot{x}$ and $\overline{x}$ *alone.* Then we add other *Z* –variables. Then we add the random effects error structure (see e.g. Cameron and Trivedi,Section 23.2.3). Finally, we look at the fixed effects estimation where *y* (not $\ddot{y}$) is regressed on $\Delta x = \ddot{x}$ only and the in-between method where *y* (not $\overline{y}$) is regressed on $\overline{x}$ *and* Z.

First, we look at the middle panel. It appears that the estimates of the time-varying explanatory variables are exactly the same under the four methods by which they are estimated. Second, we see from the first and third panel that the estimates are identical as well. The standard errors of the random effects model

[2] Discretizing into two classes only yields the hardest case. The more classes are distinguished, the more the data will look like continuously observed.

and the in-between estimation are identical. The Random Effects-standard errors are the correct ones, and as expected the Pooled OLS-standard errors differ considerably. Comparing the first and second column of Table 1 we see that adding the *Z*- variables changes the time- constant - effects , but not the effects of Δx. There are indeed individual fixed effects, which are (partly) covered by the *Z*- variables. All the results are in line with the theory developed above. Finally, comparison of the second and third panel of Table 1 shows that the coefficients of $\ddot{x}$ and $\overline{x}$ are dramatically different, while the traditional model $y = \beta' x = \beta' \ddot{x} + \beta' \overline{x}$ would suggest that they not differ statistically. Our conclusion for this data set is that the model $y = \beta' x$ is not appropriate and that $y = \beta' \ddot{x} + \beta' \overline{x}$ is the preferred alternative.

HERE TABLE 1

Probit results.

In Section 4 we argued that Probit may be applied on the latent model

$$y_{it} = \ddot{X}_{it}' \beta_{wi} + \overline{X}_i' \beta_{ib} + Z_i' \gamma + \beta_0 + \alpha_i + \varepsilon_{it} \qquad (5.1)$$

In the continuous model the presence of the fixed effect α_i gives difficulties if it is correlated with $\overline{X}_i$ and/or Z_i. The same difficulty appears in the Probit –context as well.

When we like to compare the outcomes of the OP-estimation with the OLS-results of the same data set, we have at first to create discrete events. Therefore, we discretize the data set by distinguishing the two discrete events $y_{it} \le 0, y_{it} > 0$.

The problem when applying Probit is that the error variance σ_ε in the original data set will not be equal to one, as is usually assumed when applying Probit. However, (5.1) is equivalent to

$$\frac{1}{\sigma_\varepsilon} y_{it} = \ddot{X}_{it}' \frac{\ddot{\beta}}{\sigma_\varepsilon} + \bar{X}_i' \frac{\bar{\beta}}{\sigma_\varepsilon} + Z_i' \frac{\gamma}{\sigma_\varepsilon} + \frac{\beta_0}{\sigma_\varepsilon} + \frac{\alpha_i}{\sigma_\varepsilon} + \frac{\varepsilon_{it}}{\sigma_\varepsilon} \qquad (5.2)$$

This equation has error variance equal to one. Hence, it follows that if we apply standard Probit on (5.1) we will get consistent estimators of the *ratios* between the coefficients.

Now we apply Probit on the discretized versions of the equations estimated in Table 1. This yields Table 2.

HERE TABLE 2

Comparing the ratios is facilitated by constructing Table 2A where we divided the first column in Table 2 by 0.243, the coefficient of Δ*Experience* . Similar divisions are applied to columns 2,3,4. In this way all coefficients of Δ*Experience* are set equal to one and comparison of the different methods becomes easy. This division is, of course, also applied for the standard deviations. This yields the four corresponding columns in table 2 A, which are easily compared. For the last column we multiply the cells in the last column by the factor 335/313. This yields a comparable version of the fifth column. This yields an auxiliary table 2A, where the columns are comparable, and ideally, should be equal.

HERE TABLE 2A

Taking into account the standard deviations we see that the columns are roughly equivalent and also equivalent to the results in Table 1. Looking at the second column in the second panel we find from Table 1 that the ratios are -0.004/0.116 =-0.034, and then 0.38, 0.71. The corresponding ratios in Table 2A are -0.031, 0.363, and 0.988.

Taking into account the standard deviations we see that the columns in Table2 are not significantly different and also in line with the results in Table 1.

Our main conclusion is that Probit on panel data is possible and yields consistent estimators.

6. Conclusion

In this paper we reconsidered the econometric approach to linear panel data analysis. We argue that the usual panel data model should be replaced by a model that at least includes two sets of variables: the deviations from the mean *and* the averages over time. We call this the extended model. When we do that we find that the in- between estimator and the within-estimator can be found simultaneously from the same regression equation. The two estimators appear to be uncorrelated. The pooled estimator is a weighted average of the first two estimators. By this simple extension of the model it is possible to analyze linear panels by the usual cross-section methodology. In other words, we can get rid of the rather ad hoc specifications focusing on within *or* in-between effects. In this extended model there is no need for an a priori assumption on equality of the

inter-temporal and inter- individual effects. It is rather a hypothesis which can be tested empirically. The extension of OLS to qualitatively ordered variables to be explained by means of Ordered Probit does not offer any special problem for the extended panel model. Henceforth, Ordered Probit can be used for the analysis of qualitative panel data.

The results obviously are also relevant for regional panels, and treatment evaluations (dif- in- dif studies). In this paper we looked only at one-way models (see Cameron and Trivedi ,section 21.8.) , variables are only indexed by *i* or *i,t* but generalizations to two-way models where variables are indexed by *i* or *t* or *i,t* are obvious. Also this approach seems applicable to more than two indexes.

The results of this study suggest that a special place for linear panel data analysis in econometrics as a specific subfield of econometrics is perhaps less obvious than it is thought by now.

References:

Baltagi, 2001, *Econometric Analysis of Panel Data*, 2nd edition, John Wiley & Sons Ltd, Chichester.

Cameron, C. amd P.K. Trivedi, 2005, *Microeconometrics*, Cambridge University Press, Cambridge (Mass.).

Chamberlain, G., 1984, Panel Data, Chapter 22 of the *Handbook of Econometrics*, vol. II, Z. Griliches and M.D. Intriligator (eds), Elsevier Science Publishers, Amsterdam.

Friedman,M. ,1956, *A Theory of the Consumption Function* , Princeton, NJ: Princeton University Press.

Frisch, R., Waugh, F. V. (1933). "Partial Time Regressions as Compared with Individual Trends". *Econometrica* 1 (4): 387–401. .

Hsiao, C., 1986, *Analysis of Panel Data*, Cambridge University Press, Cambridge (Mass.).

Lovell, M. (1963). "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis". *Journal of the American Statistical Association* **58** (304): 993–

Mundlak, Y., 1978, "On the Pooling of Time Series and Cross Section Data", *Econometrica*, vol. 46 (1), pp. 69-85.

Vella, F., Verbeek, M., 1998. "Whose wage do unions raise? A dynamic model of unionism and wage rate determination for young men" . *Journal of Applied Econometrics* 13, 163–183.

Verbeek, M., 2008, *A Guide to Modern Econometrics*, 3rd edition, John Wiley & Sons Ltd, Chichester.

Wooldridge, J.M., 2002, *Econometric Analysis of Cross Section and Panel Data*, MIT-Press, Cambridge (Mass.).p.487-91

Table 1: Estimation on the Vella and Verbeek (1998)-data.

Variable	OLS	OLS	Random Effects	Fixed Effects	Between
Time constant explanatory variables					
Constant		0.572	0.572		0.572
		(0.109)**	(0.224)*		(0.224)*
School		0.092	0.092		0.092
		(0.005)**	(0.011)**		(0.011)**
Black		-0.129	-0.129		-0.129
		(0.024)**	(0.049)**		(0.049)**
Hispanic		-0.034	-0.034		-0.034
		(0.022)	(0.046)		(0.046)
Time varying explanatory variables					
ΔExperience	0.116	0.116	0.116	0.116	
	(0.012)**	(0.011)**	(0.008)**	(0.008)**	
$\Delta\text{Experience}^2$	-0.004	-0.004	-0.004	-0.004	
	(0.001)**	(0.001)**	(0.001)**	(0.001)**	
ΔMarried	0.045	0.045	0.045	0.045	
	(0.026)	(0.025)	(0.018)*	(0.018)*	
ΔUnion	0.082	0.082	0.082	0.082	
	(0.027)**	(0.026)**	(0.019)**	(0.019)**	
Time-means of the time varying explanatory variables					
M: Experience	-0.023	-0.048	-0.048		-0.048
	(0.025)	(0.024)*	(0.050)		(0.050)
M: Experience^2	-0.001	0.005	0.005		0.005
	(0.002)	(0.002)**	(0.003)		(0.003)
M: Married	0.225	0.160	0.160		0.160
	(0.020)**	(0.020)**	(0.041)**		(0.041)**
M: Union	0.246	0.273	0.273		0.273
	(0.023)**	(0.023)**	(0.046)**		(0.046)**

Dependent variable: log(hourly wage). Standard errors between parenthesis. **,* = significant at 1%, 5%. Dummies for living in the North East, North Central and South (reference West, in deviation from their mean) and their means across time were also added to the specification.

Table 2: Probit estimation on the Vella and Verbeek (1998)-data.

Variable	Probit	Probit	Random Effects Probit	Fixed Effects	Between
Time constant explanatory variables					
Constant		-2.974	-4.699		-2.802
		(0.315)**	(1.008)**		(0.306)**
School		0.236	0.385		0.218
		(0.016)**	(0.050)**		(0.015)
Black		-0.335	-0.619		-0.313
		(0.069)**	(0.225)**		(0.067)**
Hispanic		-0.079	-0.183		-0.074
		(0.063)	(0.206)		(0.061)
Time varying explanatory variables					
ΔExperience	0.243	0.256	0.409	0.239	
	(0.032)**	(0.033)**	(0.042)**	(0.031)**	
$\Delta\text{Experience}^2$	-0.008	-0.008	-0.012	-0.008	
	(0.002)**	(0.002)**	(0.003)**	(0.002)**	
ΔMarried	0.082	0.093	0.154	0.081	
	(0.068)	(0.070)	(0.092)	(0.067)	
ΔUnion	0.244	0.253	0.407	0.235	
	(0.071)**	(0.073)**	(0.093)**	(0.070)**	
Time-means of the time varying explanatory variables					
M: Experience	0.051	-0.026	-0.110		-0.007
	(0.070)	(0.072)	(0.223)		(0.070)
M: Experience^2	-0.010	0.005	0.013		0.003
	(0.004)*	(0.005)	(0.014)		(0.005)
M: Married	0.476	0.332	0.575		0.315
	(0.054)**	(0.056)**	(0.185)**		(0.055)**
M: Union	0.688	0.790	1.357		0.743
	(0.063)**	(0.065)**	(0.215)**		(0.064)**

Dependent variable: log(hourly wage). Standard errors between parenthesis. The standard errors reported for the Pooled OLS-estimation are OLS- and clustered across individuals-standard errors. **,* = significant at 1%, 5%. Dummies for living in the North East, North Central and South (reference West, in deviation from their mean) and their means across time were also added to the specification.

Table 2A: Probit estimates normalized.

Variable	Probit	Probit	Random Effects Probit	Dummy	Dummy
Time constant explanatory variables					
Constant		-11.62	-11.49		-11.72
		(1.230)**	(2.465)**		(1.280)**
School		0.922	0.941		0.912
		(0.062)**	(0.122)**		(0.063)
Black		-1.309	-1.513		-1.309
		(0.270)**	(0.550)**		(0.280)**
Hispanic		-0.309	-0.447		-0.309
		(0.246)	(0.504)		(0.255)
Time varying explanatory variables					
ΔExperience	1.000	1.000	1.000	1.000	
	(0.132)**	(0.129)**	(0.103)**	(0.130)**	
$\Delta Experience^2$	-0.033	-0.031	-0.029	-0.033	
	(0.008)**	(0.008)**	(0.007)**	(0.008)**	
ΔMarried	0.337	0.363	0.377	0.339	
	(0.280)	(0.273)	(0.225)	(0.280)	
ΔUnion	1.004	0.988	0.995	0.983	
	(0.292)**	(0.285)**	(0.227)**	(0.293)**	
Time-means of the time varying explanatory variables					
M: Experience	0.210	-0.102	-0.269		-0.029
	(0.288)	(0.281)	(0.545)		(0.293)
M: $Experience^2$	-0.041	0.020	0.032		0.013
	(0.016)*	(0.020)	(0.034)		(0.021)
M: Married	1.959	1.297	1.406		1.317
	(0.222)**	(0.219)**	(0.452)**		(0.230)**
M: Union	2.831	3.086	3.318		3.107
	(0.259)**	(0.254)**	(0.526)**		(0.268)**