# Static Panel Data Models[1]

**Class Notes**

Manuel Arellano

Revised: October 9, 2009

## 1 Unobserved Heterogeneity

The econometric interest in panel data has been the result of two different motivations:

- First, the desire of exploiting panel data for controlling unobserved time-invariant heterogeneity in cross-sectional models.
- Second, the use of panel data to disentangle components of variance and estimating transition probabilities, and more generally to study the dynamics of cross-sectional populations.

These motivations can be loosely associated with two strands of the panel data literature labelled *fixed effects* and *random effects* models. We next take these two motivations and models in turn.

### 1.1 Overview

A sizeable part of econometric activity deals with empirical description and forecasting, but another aims at quantifying structural or causal relationships. Structural relations are needed for policy evaluation and for testing theories.

The regression model is an essential tool for descriptive and structural econometrics. However, regression lines from economic data often cannot be given a causal interpretation. The reason is that sometimes we expect correlation between explanatory variables and errors in the relation of interest.

One example is the classical supply-and-demand simultaneity problem due market equilibrium. Another is measurement error: if the explanatory variable we observe is not the variable to whom agents respond but an error ridden measure of it, the unobservable term in the equation of interest will contain the measurement error, which will be correlated with the regressor. Finally, there may be correlation due to unobserved heterogeneity. This has been a pervasive problem in cross-sectional regression analysis. If characteristics that have a direct effect on both left- and right-hand side variables are omitted, explanatory variables will be correlated with errors and regression coefficients will be biased measures of the structural effects. Thus, researchers have often been confronted with massive cross-sectional data sets from which precise correlations can be determined but that, nevertheless, had no information about parameters of policy interest.

[1] This is an abridged version of Part I in Arellano (2003).

The traditional response of econometrics to these problems has been multiple regression and instrumental variable models. Regrettably, we often lack data on the conditioning variables or the instruments to achieve identification of structural parameters in these ways.

A major motivation for using panel data has been the ability to control for possibly correlated, time-invariant heterogeneity without observing it. Suppose a cross-sectional regression of the form

$$y_{i1} = \beta x_{i1} + \eta_i + v_{i1} \tag{1}$$

such that $E(v_{i1} \mid x_{i1}, \eta_i) = 0$. If $\eta_i$ is observed $\beta$ can be identified from a multiple regression of $y$ on $x$ and $\eta$. If $\eta_i$ is not observed identification of $\beta$ requires either lack of correlation between $x_{i1}$ and $\eta_i$, in which case

$$Cov(x_{i1,}\eta_i) = 0 \Rightarrow \beta = \frac{Cov(x_{i1}, y_{i1})}{Var(x_{i1})},$$

or the availability of an external instrument $z_i$ that is uncorrelated with both $v_{i1}$ and $\eta_i$ but correlated with $x_{i1}$, in which case

$$Cov(z_{i,}\eta_i) = 0 \Rightarrow \beta = \frac{Cov(z_i, y_{i1})}{Cov(z_i, x_{i1})}.$$

Suppose that neither of these two options is available, but we observe $y_{i2}$ and $x_{i2}$ for the same individuals in a second period (so that $T = 2$) such that

$$y_{i2} = \beta x_{i2} + \eta_i + v_{i2} \tag{2}$$

and both $v_{i1}$ and $v_{i2}$ satisfy $E(v_{it} \mid x_{i1}, x_{i2}, \eta_i) = 0$. Then $\beta$ is identified in the regression in first-differences even if $\eta_i$ is not observed. We have:

$$y_{i2} - y_{i1} = \beta(x_{i2} - x_{i1}) + (v_{i2} - v_{i1}) \tag{3}$$

and

$$\beta = \frac{Cov(\Delta x_{i2}, \Delta y_{i2})}{Var(\Delta x_{i2})}. \tag{4}$$

**A Classic Example: Agricultural Production (Mundlak 1961, Chamberlain 1984)** Suppose equation (1) represents the Cobb-Douglas production function of an agricultural product. The index $i$ denotes farms and $t$ time periods (seasons or years). Also:

$y_{it}$ = Log output.

$x_{it}$ = Log of a variable input (labour).

$\eta_i$ = An input that remains constant over time (soil quality).

$v_{it}$ = A stochastic input which is outside the farmer's control (rainfall).

Suppose $\eta_i$ is known by the farmer but not by the econometrician. If farmers maximize expected profits there will be cross-sectional correlation between labour and soil quality. Therefore, the population coefficient in a simple regression of $y_{i1}$ on $x_{i1}$ will differ from $\beta$. If $\eta$ were observed by the econometrician, the coefficient on $x$ in a multiple cross-sectional regression of $y_{i1}$ on $x_{i1}$ and $\eta_i$ will coincide with $\beta$. Now suppose that data on $y_{i2}$ and $x_{i2}$ for a second period become available. Moreover, suppose that rainfall in the second period is unpredictable from rainfall in the first period (permanent differences in rainfall are part of $\eta_i$), so that rainfall is independent of a farm's labour demand in the two periods. Thus, even in the absence of data on $\eta_i$ the availability of panel data affords the identification of the technological parameter $\beta$.

**A Firm Money Demand Example (Bover and Watson, 2005)** Suppose firms minimize cost for given output $s_{it}$ subject to a production function $s_{it} = F(x_{it})$ and to some transaction services $s_{it} = \left(a_i m_{it}^{(1-b)} \ell_{it}^b\right)^{1/c}$, where $x$ denotes a composite input, $m$ is demand for cash, $\ell$ is labour employed in transactions, and $a$ represents the firm's financial sophistication. There will be economies of scale in the demand for money by firms if $c \neq 1$. The resulting money demand equation is

$$\log m_{it} = k + c \log s_{it} - b \log(R_{it}/w_{it}) - \log a_i + v_{it}. \tag{5}$$

Here $k$ is a constant, $R$ is the opportunity cost of holding money, $w$ is the wage of workers involved in transaction services, and $v$ is a measurement error in the demand for cash. In general $a$ will be correlated with output through the cash-in-advance constraint. Thus, the coefficient of output (or sales) in a regression of $\log m$ on $\log s$ and $\log(R/w)$ will not coincide with the scale parameter of interest. However, if firm panel data is available and $a$ varies across firms but not over time in the period of analysis, economies of scale can be identified from the regression in changes.

**An Example in which Panel Data Does Not Work: Returns to Education** "Structural" returns to education are important in the assessment of educational policies. It has been widely believed in the literature that cross-sectional regression estimates of the returns could not be trusted because of omitted "ability" potentially correlated with education attainment. In the earlier notation:

$y_{it}$ = Log wage (or earnings).

$x_{it}$ = Years of full-time education.

$\eta_i$ = Unobserved ability.

$\beta$ = Returns to education.

The problem in this example is that $x_{it}$ typically lacks time variation. So a regression in first-differences will not be able to identify $\beta$ in this case. In this context data on siblings and cross-sectional instrumental variables have proved more useful for identifying returns to schooling free of ability bias than panel data (Griliches, 1977).

This example illustrates a more general problem. Information about $\beta$ in the regression in first-differences will depend on the ratio of the variances of $\Delta v$ and $\Delta x$. In the earnings$-$education equation, we are in the extreme situation where $Var(\Delta x) = 0$, but if $Var(\Delta x)$ is small regressions in changes may contain very little information about parameters of interest even if the cross-sectional sample size is very large.

**Econometric Measurement *versus* Forecasting Problems** The previous examples suggest that the ability to control for unobserved heterogeneity is mainly an advantage in the context of problems of econometric measurement as opposed to problems of forecasting. This is an important distinction. Including individual effects we manage to identify certain coefficients at the expense of leaving part of the regression unmodelled (the one that only has cross-sectional variation).

The part of the variance of $y$ accounted by $x\beta$ could be very small relative to $\eta$ and $v$. In a case like this it would be easy to obtain higher $R^2$ by including lagged dependent variables or proxies for the fixed effects. Regressions of this type would be useful in cross-sectional forecasting exercises for the population from which the data come (like in credit scoring or in the estimation of probabilities of tax fraud), but they may be of no use if the objective is to measure the effect of $x$ on $y$ holding constant all time-invariant heterogeneity.

An equation with individual specific intercepts may still be useful when the interest is in forecasts for the same individuals in different time periods, but not when we are interested in forecasts for individuals other than those included in the sample.

**Non-Exogeneity and Random Coefficients** The identification of causal effects through regression coefficients in differences or deviations depends on the lack of correlation between $x$ and $v$ at all lags and leads (strict exogeneity). If $x$ is measured with error or is correlated with lagged errors, regressions in deviations may actually make things worse.

Another difficulty arises when the effect of $x$ on $y$ is itself heterogeneous. In such case regression coefficients in differences cannot in general be interpreted as average causal effects. Specifically, suppose that $\beta$ is allowed to vary cross-sectionally in (1) and (2) so that

$$y_{it} = \beta_i x_{it} + \eta_i + v_{it} \;\; (t = 1, 2) \;\; E\left(v_{it} \mid x_{i1}, x_{i2}, \eta_i, \beta_i\right) = 0. \tag{6}$$

In these circumstances, the regression coefficient (4) differs from $E\left(\beta_i\right)$ unless $\beta_i$ is mean independent of $\Delta x_{i2}$. The availability of panel data still affords identification of average causal effects in random coefficients models as long as $x$ is strictly exogenous. However, if $x$ is not exogenous and $\beta_i$ is heterogeneous we run into serious identification problems in short panels.

## 1.2 Fixed Effects Models

### 1.2.1 Assumptions

Our basic assumptions for what we call the "static fixed effects model" are as follows. We assume that $\{(y_{i1}, ..., y_{iT}, x_{i1}, ..., x_{iT}, \eta_i), i = 1, ..., N\}$ is a random sample and that

$$y_{it} = x'_{it}\beta + \eta_i + v_{it} \tag{7}$$

together with

*Assumption A1:*

$$E(v_i \mid x_i, \eta_i) = 0 \ (t = 1, ..., T),$$

where $v_i = (v_{i1}, ..., v_{iT})'$ and $x_i = (x'_{i1}, ..., x'_{iT})'$. We observe $y_{it}$ and the $k \times 1$ vector of explanatory variables $x_{it}$ but not $\eta_i$, which is therefore an unobservable time-invariant regressor.

Similarly, we shall refer to "classical" errors when the additional auxiliary assumption holds:

*Assumption A2:*

$$Var(v_i \mid x_i, \eta_i) = \sigma^2 I_T.$$

Under Assumption *A2* the errors are conditionally homoskedastic and not serially correlated.

Under Assumption *A1* we have

$$E(y_i \mid x_i, \eta_i) = X_i\beta + \eta_i\iota \tag{8}$$

where $y_i = (y_{i1}, ..., y_{iT})'$, $\iota$ is a $T \times 1$ vector of ones, and $X_i = (x_{i1}, ..., x_{iT})'$ is a $T \times k$ matrix. The implication of (8) for the expected value of $y_i$ given $x_i$ is

$$E(y_i \mid x_i) = X_i\beta + E(\eta_i \mid x_i)\iota. \tag{9}$$

Moreover, under Assumption *A2*

$$Var(y_i \mid x_i, \eta_i) = \sigma^2 I_T, \tag{10}$$

which implies

$$Var(y_i \mid x_i) = \sigma^2 I_T + Var(\eta_i \mid x_i)\iota\iota'. \tag{11}$$

*A1* is the fundamental assumption in this context. It implies that the error $v$ at any period is uncorrelated with past, present, and future values of $x$ (or, conversely, that $x$ at any period is uncorrelated with past, present, and future values of $v$). *A1* is, therefore, an assumption of *strict exogeneity* that rules out, for example, the possibility that current values of $x$ are influenced by past errors. In the agricultural production function example, $x$ (labour) will be uncorrelated with $v$ (rainfall) at all lags and leads provided the latter is unpredictable from past rainfall (given permanent differences in rainfall that would be subsumed in the farm effects, and possibly seasonal or other deterministic components). If rainfall in period $t$ is predictable from rainfall in period $t-1$ —which is known to the farmer in $t$—labour demand in period $t$ will in general depend on $v_{i(t-1)}$ (Chamberlain, 1984, 1258–1259).

Assumption *A2* is, on the other hand, an auxiliary assumption under which classical least-squares results are optimal. However, lack of compliance with *A2* is often to be expected in applications. Here, we first present results under *A2,* and subsequently discuss estimation and inference with heteroskedastic and serially correlated errors.

As for the nature of the effects, strictly speaking, the term *fixed effects* would refer to a sampling process in which the same units are (possibly) repeatedly sampled for a given period holding constant the effects. In such context one often has in mind a distribution of individual effects chosen by the researcher. Here we imagine a sample randomly drawn from a multivariate population of observable data and unobservable effects. This notion may or may not correspond to the physical nature of data collection. It would be so, for example, in the case of some household surveys, but not with data on all quoted firms or OECD countries. In those cases, the multivariate population from which the data come is hypothetical. Moreover, we are interested in models which only specify features of the conditional distribution $f(y_i \mid x_i, \eta_i)$. Therefore, we are not concerned with whether the distribution that generates the data on $x_i$ and $\eta_i$, $f(x_i, \eta_i)$ say, is representative of some cross-sectional population or of the researcher's wishes. We just regard $(y_i, x_i, \eta_i)$ as a random sample from the (perhaps artificial) multivariate population with joint distribution $f(y_i, x_i, \eta_i) = f(y_i \mid x_i, \eta_i) f(x_i, \eta_i)$ and focus on the conditional distribution of $y_i$. So in common with much of the econometric literature, we use the term fixed effects to refer to a situation in which $f(\eta_i \mid x_i)$ is left unrestricted.

### 1.2.2 Within-Group Estimation

With $T = 2$ there is just one equation after differencing. Under Assumptions *A1* and *A2*, the equation in first differences is a classical regression model and hence OLS in first-differences is the optimal estimator of $\beta$ in the least squares sense. To see the irrelevance of the equations in levels in this model, note that a non-singular transformation of the original two-equation system is

$$E(y_{i1} \mid x_i) = x_{i1}'\beta + E(\eta_i \mid x_i)$$

$$E(\Delta y_{i2} \mid x_i) = \Delta x_{i2}'\beta.$$

Since $E(\eta_i \mid x_i)$ is an unknown unrestricted function of $x_i$, knowledge of the function $E(y_{i1} \mid x_i)$ is uninformative about $\beta$ in the first-equation. Thus, no information about $\beta$ is lost by only considering the equation in first-differences.

If $T \geq 3$ we have a system of $T-1$ equations in first-differences:

$$\begin{aligned}
\Delta y_{i2} &= \Delta x_{i2}'\beta + \Delta v_{i2} \\
&\vdots \\
\Delta y_{iT} &= \Delta x_{iT}'\beta + \Delta v_{iT},
\end{aligned}$$

which in compact form can be written as

$$Dy_i = DX_i\beta + Dv_i, \tag{12}$$

where $D$ is the $(T-1) \times T$ matrix first-difference operator

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \tag{13}$$

Provided each of the errors in first-differences are mean independent of the $x$s for all periods (under Assumption *A1*) $E(Dv_i \mid x_i) = 0$, OLS estimates of $\beta$ in this system given by

$$\widehat{\beta}_{OLS} = \left( \sum_{i=1}^{N} (DX_i)' \, DX_i \right)^{-1} \sum_{i=1}^{N} (DX_i)' \, Dy_i \tag{14}$$

will be unbiased and consistent for large $N$. However, if the $v$s are homoskedastic and non-autocorrelated classical errors (under Assumption *A2*), the errors in first-differences will be correlated for adjacent periods with

$$Var(Dv_i \mid x_i) = \sigma^2 DD'. \tag{15}$$

Following standard regression theory, the optimal estimator in this case is given by generalized least-squares (GLS), which takes the form

$$\widehat{\beta}_{WG} = \left( \sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} DX_i \right)^{-1} \sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} Dy_i. \tag{16}$$

In this case GLS itself is a feasible estimator since $DD'$ does not depend on unknown coefficients.

The idempotent matrix $D' \left(DD'\right)^{-1} D$ also takes the form[2]

$$D' \left(DD'\right)^{-1} D = I_T - \iota\iota'/T \equiv Q, \text{ say.} \tag{17}$$

[2] To verify this, note that the $T \times T$ matrix

$$\mathcal{H} = \begin{pmatrix} T^{-1/2}\iota' \\ \left(DD'\right)^{-1/2} D \end{pmatrix}$$

is such that $\mathcal{H}\mathcal{H}' = I_T$, so that also $\mathcal{H}'\mathcal{H} = I_T$ or

$$\iota\iota'/T + D' \left(DD'\right)^{-1} D = I_T.$$

The matrix $Q$ is known as the *within-group* operator because it transforms the original time series into deviations from time means: $\widetilde{y}_i = Qy_i$, whose elements are given by

$$\widetilde{y}_{it} = y_{it} - \overline{y}_i$$

with $\overline{y}_i = T^{-1}\sum_{s=1}^{T} y_{is}$. Therefore, $\widehat{\beta}_{WG}$ can also be expressed as OLS in deviations from time means

$$\widehat{\beta}_{WG} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)(x_{it}-\overline{x}_i)'\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)(y_{it}-\overline{y}_i). \tag{18}$$

This is probably the most popular estimator in panel data analysis, and it is known under a variety of names including within-group and covariance estimator.[3]

It is also known as the *dummy-variable least-squares* or "fixed effects" estimator. This name reflects the fact that since $\widehat{\beta}_{WG}$ is a least-squares estimator after subtracting individual means to the observations, it is numerically the same as the estimator of $\beta$ that would be obtained in a OLS regression of $y$ on $x$ and a set of $N$ dummy variables, one for each individual in the sample. Thus $\widehat{\beta}_{WG}$ can also be regarded as the result of estimating jointly by OLS $\beta$ and the realizations of the individual effects that appear in the sample.

To see this, consider the system of $T$ equations in levels

$$y_i = X_i\beta + \iota\eta_i + v_i$$

and write it in stacked form as

$$y = X\beta + C\eta + v, \tag{19}$$

where $y = (y_1', ..., y_N')'$ and $v = (v_1', ..., v_N')'$ are $NT \times 1$ vectors, $X = (X_1', ..., X_N')'$ is an $NT \times k$ matrix, $C$ is an $NT \times N$ matrix of individual dummies given by $C = I_N \otimes \iota$, and $\eta = (\eta_1, ..., \eta_N)'$ is the $N \times 1$ vector of individual specific effects or intercepts. Using the result from partitioned regression, the OLS regression of $y$ on $X$ and $C$ gives the following expression for estimated $\beta$

$$\left[X'\left(I_{NT} - C(C'C)^{-1}C'\right)X\right]^{-1} X'\left(I_{NT} - C(C'C)^{-1}C'\right)y, \tag{20}$$

which clearly coincides with $\widehat{\beta}_{WG}$ since $I_{NT} - C(C'C)^{-1}C' = I_N \otimes Q$.

The expressions for the estimated effects are

$$\widehat{\eta}_i = \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\beta}_{WG}\right) \equiv \overline{y}_i - \overline{x}_i'\widehat{\beta}_{WG}\ (i = 1, ..., N). \tag{21}$$

[3] The name "within-group" originated in the context of data with a group structure (like data on families and family members). Panel data can be regarded as a special case of this type of data in which the "group" is formed by the time series observations from a given individual.

We do not need to go beyond standard regression theory to obtain the sampling properties of these estimators. The fact that $\widehat{\beta}_{WG}$ is the GLS for the system of $T-1$ equations in first-differences tells us that it will be unbiased and optimal in finite samples. It will also be consistent as $N$ tends to infinity for fixed $T$ and asymptotically normal under usual regularity conditions. The $\widehat{\eta}_i$ will also be unbiased estimates of the $\eta_i$ for samples of any size, but being time series averages, their variance can only tend to zero as $T$ tends to infinity. Therefore, they cannot be consistent estimates for fixed $T$ and large $N$. Clearly, the within-group estimates $\widehat{\beta}_{WG}$ will also be consistent as $T$ tends to infinity regardless of whether $N$ is fixed or not.

Fixed effects models have a long tradition in econometrics. Their use was first suggested in two Cowles Commission papers by Clifford Hildreth in 1949 and 1950, and early applications were conducted by Mundlak (1961) and Hoch (1962). The motivation in these two studies was to rely on fixed effects in order to control for simultaneity bias in the estimation of agricultural production functions.

**Orthogonal Deviations** Finally, it is worth finding out the form of the transformation to the original data that results from doing first-differences and further applying a GLS transformation to the differenced data to remove the moving-average serial correlation induced by differencing (Arellano and Bover, 1995). The required transformation is given by the $(T-1) \times T$ matrix

$$A = \left(DD'\right)^{-1/2} D.$$

If we choose $(DD')^{-1/2}$ to be the upper triangular Cholesky factorization, the operator $A$ is such that a $T \times 1$ time series error transformed by $A$, $v_i^* = Av_i$ consists of $T-1$ elements of the form

$$v_{it}^* = c_t[v_{it} - \frac{1}{(T-t)}(v_{i(t+1)} + ...v_{iT})] \tag{22}$$

where $c_t^2 = (T-t)/(T-t+1)$. Clearly, $A'A = Q$ and $AA' = I_{T-1}$. We then refer to this transformation as *forward orthogonal deviations*. Thus, if $Var(v_i) = \sigma^2 I_T$ we also have $Var(v_i^*) = \sigma^2 I_{T-1}$. So orthogonal deviations can be regarded as an alternative transformation, which in common with first-differencing eliminates individual effects but in contrast it does not introduce serial correlation in the transformed errors. Moreover, the within-group estimator can also be regarded as OLS in orthogonal deviations.

## 1.3 Heteroskedasticity and Serial Correlation

### 1.3.1 Robust Standard Errors for Within-Group Estimators

If assumption *A1* holds but *A2* does not (that is, using orthogonal deviations, if $E(v_i^* \mid x_i) = 0$ but $Var(v_i^* \mid x_i) \neq \sigma^2 I_{T-1}$), the ordinary regression formulae for estimating the within-group variance will lead to inconsistent standard errors. Such formula is given by

$$\widehat{Var}(\widehat{\beta}_{WG}) = \widehat{\sigma}^2 (X^{*\prime} X^*)^{-1} \tag{23}$$

where $X^* = (I_N \otimes A)X$, $y^* = (I_N \otimes A)y$, and $\widehat{\sigma}^2$ is the unbiased residual variance

$$\widehat{\sigma}^2 = \frac{1}{N(T-1)-k}(y^* - X^*\widehat{\beta}_{WG})'(y^* - X^*\widehat{\beta}_{WG}). \tag{24}$$

However, since

$$\left(\frac{1}{N}X^{*\prime}X^*\right)\sqrt{N}(\widehat{\beta}_{WG} - \beta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i^{*\prime}v_i^*$$

and $E(X_i^{*\prime}v_i^*) = 0$, the right-hand side of the previous expression is a scaled sample average of zero-mean random variables to which a standard central limit theorem for multivariate *iid* observations can be applied for fixed $T$ as $N$ tends to infinity:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i^{*\prime}v_i^* \xrightarrow{d} \mathcal{N}\left[0, E(X_i^{*\prime}v_i^*v_i^{*\prime}X_i^*)\right].$$

Therefore, an estimate of the asymptotic variance of the within-group estimator that is robust to heteroskedasticity and serial correlation of arbitrary forms for fixed $T$ and large $N$ can be obtained as

$$\widetilde{Var}(\widehat{\beta}_{WG}) = (X^{*\prime}X^*)^{-1}\left(\sum_{i=1}^{N} X_i^{*\prime}\widehat{v}_i^*\widehat{v}_i^{*\prime}X_i^*\right)(X^{*\prime}X^*)^{-1} \tag{25}$$

with $\widehat{v}_i^* = y_i^* - X_i^*\widehat{\beta}_{WG}$ (Arellano, 1987). The square root of diagonal elements of $\widetilde{Var}(\widehat{\beta}_{WG})$ provide standard errors clustered by individual.

### 1.3.2 Optimal GLS with Heteroskedasticity and Autocorrelation of Unknown Form

If $Var(v_i^* \mid x_i) = \Omega(x_i)$ where $\Omega(x_i)$ is a symmetric matrix of order $T-1$ containing unknown functions of $x_i$, the optimal estimator of $\beta$ will be of the form

$$\widehat{\beta}_{UGLS} = \left(\sum_{i=1}^{N} X_i^{*\prime}\Omega^{-1}(x_i)X_i^*\right)^{-1}\sum_{i=1}^{N} X_i^{*\prime}\Omega^{-1}(x_i)y_i^*. \tag{26}$$

This estimator is unfeasible because $\Omega(x_i)$ is unknown. A feasible semi parametric GLS estimator would use instead a nonparametric estimator of $E(v_i^*v_i^{*\prime} \mid x_i)$ based on within-group residuals. Under appropriate regularity conditions and a suitable choice of nonparametric estimator, feasible GLS can be shown to attain for large $N$ the same efficiency as $\widehat{\beta}_{UGLS}$.

A special case which gives rise to a straightforward feasible GLS (for small $T$ and large $N$), first discussed by Kiefer (1980), is one in which the conditional variance of $v_i^*$ is a constant but non-scalar matrix: $Var(v_i^* \mid x_i) = \Omega$. This assumption rules out conditional heteroskedasticity, but allows for autocorrelation and unconditional time series heteroskedasticity in the original equation errors $v_{it}$. In this case, a feasible GLS estimator takes the form

$$\widehat{\beta}_{FGLS} = \left(\sum_{i=1}^{N} X_i^{*\prime}\widehat{\Omega}^{-1}X_i^*\right)^{-1}\sum_{i=1}^{N} X_i^{*\prime}\widehat{\Omega}^{-1}y_i^* \tag{27}$$

where $\widehat{\Omega}$ is given by the orthogonal-deviation WG residual intertemporal covariance matrix

$$\widehat{\Omega} = \frac{1}{N} \sum_{i=1}^{N} \widehat{v}_i^* \widehat{v}_i^{*\prime}. \tag{28}$$

### 1.3.3 Improved GMM under Heteroskedasticity and Autocorrelation of Unknown Form

The basic condition $E(v_i^* \mid x_i) = 0$ implies that any function of $x_i$ is uncorrelated to $v_i^*$ and therefore a potential instrumental variable. Thus, any list of moment conditions of the form

$$E\left[h_t(x_i) v_{it}^*\right] = 0 \ (t = 1, ..., T-1) \tag{29}$$

for given functions $h_t(x_i)$ such that $\beta$ is identified from (29), could be used to obtain a consistent GMM estimator of $\beta$.

Under $\Omega(x_i) = \sigma^2 I_{T-1}$ the optimal moment conditions are given by

$$E\left(X_i^{*\prime} v_i^*\right) = 0, \tag{30}$$

in the sense that the variance of the corresponding optimal method-of-moments estimator (which in this case is OLS in orthogonal deviations, or the WG estimator) cannot be reduced by using other functions of $x_i$ as instruments in addition to (30).

For arbitrary $\Omega(x_i)$ the optimal moment conditions are

$$E\left[X_i^{*\prime} \Omega^{-1}(x_i) v_i^*\right] = 0, \tag{31}$$

which gives rise to the optimal GLS estimator $\widehat{\beta}_{UGLS}$ given in (26).

The $k$ moment conditions (31), however, cannot be directly used because $\Omega(x_i)$ is unknown. The simpler, improved estimators that we consider in this section are based on the fact that optimal GMM from a wider list of moments than (30) can be asymptotically more efficient than WG when $\Omega(x_i) \neq \sigma^2 I_{T-1}$, although not as efficient as optimal GLS. In particular, it seems natural to consider GMM estimators of the system of $T-1$ equations in orthogonal deviations (or first-differences) using the explanatory variables for all time periods as separate instruments for each equation:

$$E\left(v_i^* \otimes x_i\right) = 0. \tag{32}$$

Note that the $k$ moments in (30) are linear combinations of the much larger set of $kT(T-1)$ moments contained in (32). Also, it is convenient to write (32) as

$$E\left(Z_i' v_i^*\right) \equiv E\left[Z_i'(y_i^* - X_i^* \beta)\right] = 0 \tag{33}$$

where $Z_i = (I_{T-1} \otimes x_i')$. With this notation, the optimal GMM estimator from (32) or (33) is given by

$$\widehat{\beta}_{GMM} = \left[\left(\sum_i X_i^{*\prime} Z_i\right) A_N \left(\sum_i Z_i' X_i^*\right)\right]^{-1} \left(\sum_i X_i^{*\prime} Z_i\right) A_N \left(\sum_i Z_i' y_i^*\right). \tag{34}$$

Optimality requires that the weight matrix $A_N$ is a consistent estimate up to a multiplicative constant of the inverse of the variance of the orthogonality conditions $E\left(Z_i' v_i^* v_i^{*\prime} Z_i\right)$.

Under Assumption *A2* $E\left(Z_i' v_i^* v_i^{*\prime} Z_i\right) = \sigma^2 E\left(Z_i' Z_i\right)$, and therefore an optimal choice is $A_N = \left(\sum_i Z_i' Z_i\right)^{-1}$. In such a case the resulting estimator is numerically the same as the within-group estimator because the columns in $X_i^*$ are linear combinations of those in $Z_i$.

More generally, an optimal choice under heteroskedasticity and serial correlation of unknown form is given by

$$A_N = \left(\sum_i Z_i' \widehat{v}_i^* \widehat{v}_i^{*\prime} Z_i\right)^{-1}. \tag{35}$$

The resulting estimator, $\widehat{\beta}_{OGMM}$ say, will be asymptotically equivalent to WG under Assumption *A2* but strictly more efficient for large $N$ when the assumption is violated. It will, nevertheless, be inefficient relative to $\widehat{\beta}_{UGLS}$. The relationship among the large sample variances of the three estimators is therefore

$$Var(\widehat{\beta}_{UGLS}) \leq Var(\widehat{\beta}_{OGMM}) \leq Var(\widehat{\beta}_{WG}),$$

with equality in both cases when Assumption *A2* holds.

Estimators of the previous type were considered by Chamberlain (1982, 1984) who motivated them as minimum distance estimators from a linear projection of $y_i$ on $x_i$ (the "Π matrix" approach).

## 1.4 Likelihood Approaches

The within-group estimator can be regarded as the Gaussian maximum likelihood estimator under three different likelihood approaches—joint, conditional, and marginal—relative to the individual effects. This is a special feature of the static linear model. In other models, different likelihood approaches give rise to different estimators. Nevertheless, regardless of their maxima, the alternative likelihood functions for the static model that we discuss in this section may be of interest in their own right from a Bayesian perspective.

### 1.4.1 Joint Likelihood

Under the normality assumption:

$$y_i \mid x_i, \eta_i \sim \mathcal{N}\left(X_i\beta + \eta_i\iota, \sigma^2 I_T\right), \tag{36}$$

the log conditional density of $y_i$ given $x_i$ and $\eta_i$ takes the form

$$\log f\left(y_i \mid x_i, \eta_i\right) \propto -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2} v_i{}' v_i \tag{37}$$

where $v_i = (y_i - X_i\beta - \eta_i\iota)$. Thus, the log likelihood of a cross-sectional sample of independent observations is a function of $\beta$, $\sigma^2$, and $\eta_1, ..., \eta_N$:

$$L\left(\beta, \sigma^2, \eta; y, x\right) = \sum_{i=1}^{N} \log f\left(y_i \mid x_i, \eta_i\right). \tag{38}$$

In view of our previous discussion and standard linear regression maximum likelihood (ML) estimation, joint maximization of (38) with respect to $\beta$, $\eta$, and $\sigma^2$ yields the WG estimator for $\beta$, the residual estimates for $\eta$ given in (21), and the residual variance without degrees of freedom correction for $\sigma^2$:

$$\widetilde{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N} \widehat{v}_i{}'\widehat{v}_i \tag{39}$$

where $\widehat{v}_i = \left(y_i - X_i\widehat{\beta}_{WG} - \widehat{\eta}_i\iota\right)$.

Unlike (24) $\widetilde{\sigma}^2$ will not be a consistent estimator of $\sigma^2$ for large $N$ and small $T$ panels. In effect, since $E\left(\sum_{i=1}^{N} \widehat{v}_i{}'\widehat{v}_i\right) = (NT - N - k)\,\sigma^2$, we have

$$\operatorname*{plim}_{N\to\infty} \widetilde{\sigma}^2 = \frac{(T-1)}{T}\sigma^2.$$

Thus $\widetilde{\sigma}^2$ has a negative (cross-sectional) large sample bias given by $\sigma^2/T$. This is an example of the *incidental parameter problem* studied by Neyman and Scott (1948). The problem is that the maximum likelihood estimator need not be consistent when the likelihood depends on a subset of (incidental) parameters whose number increases with sample size. In our case, the likelihood depends on $\beta$, $\sigma^2$ and the incidental parameters $\eta_1, ..., \eta_N$. The ML estimator of $\beta$ is consistent but that of $\sigma^2$ is not.

### 1.4.2 Conditional Likelihood

In the linear static model, $\overline{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for $\eta_i$. This means that the density of $y_i$ given $x_i$, $\eta_i$, and $\overline{y}_i$ does not depend on $\eta_i$

$$f\left(y_i \mid x_i, \eta_i, \overline{y}_i\right) = f\left(y_i \mid x_i, \overline{y}_i\right). \tag{40}$$

To see this, note that, expressing the conditional density of $y_i$ given $\overline{y}_i$ as a ratio of the joint and marginal densities, we have

$$f\left(y_i \mid x_i, \eta_i, \overline{y}_i\right) = \frac{f\left(y_i \mid x_i, \eta_i\right)}{f\left(\overline{y}_i \mid x_i, \eta_i\right)}$$

and that under (36)

$$\overline{y}_i \mid x_i, \eta_i \sim \mathcal{N}\left(\overline{x}_i'\beta + \eta_i, \frac{\sigma^2}{T}\right),$$

so that

$$\log f\left(\overline{y}_i \mid x_i, \eta_i\right) \propto -\frac{1}{2}\log\sigma^2 - \frac{T}{2\sigma^2}\overline{v}_i^2. \tag{41}$$

Subtracting (41) from (37) we obtain:

$$\log f\left(y_i \mid x_i, \eta_i, \overline{y}_i\right) \propto -\frac{(T-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(v_{it} - \overline{v}_i\right)^2, \tag{42}$$

which does not depend on $\eta_i$ because it is only a function of the within-group errors.

Thus the conditional log likelihood

$$L_c\left(\beta, \sigma^2; y, x\right) = \sum_{i=1}^{N}\log f\left(y_i \mid x_i, \overline{y}_i\right) \tag{43}$$

is a function of $\beta$ and $\sigma^2$, which can be used as an alternative basis for inference. The maximizers of (43) are the WG estimator of $\beta$ and

$$\overline{\sigma}^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}\widehat{v}_i{}'\widehat{v}_i. \tag{44}$$

Note that contrary to (39), (44) is consistent for large $N$ and small $T$, although it is not exactly unbiased as (24).

### 1.4.3 Marginal (or Integrated) Likelihood

Finally, we may consider the marginal distribution of $y_i$ given $x_i$ but not $\eta_i$:

$$f\left(y_i \mid x_i\right) = \int f\left(y_i \mid x_i, \eta_i\right) dF\left(\eta_i \mid x_i\right)$$

where $F\left(\eta_i \mid x_i\right)$ denotes the conditional *cdf* of $\eta_i$ given $x_i$. One possibility, in the spirit of the GMM approach discussed in the previous section, is to assume

$$\eta_i \mid x_i \sim \mathcal{N}\left(\delta + \lambda' x_i, \sigma_\eta^2\right), \tag{45}$$

but it is of some interest to study the form of $f\left(y_i \mid x_i\right)$ for arbitrary $F\left(\eta_i \mid x_i\right)$.

Let us consider the non-singular transformation matrix

$$H = \begin{pmatrix} T^{-1}\iota' \\ A \end{pmatrix}. \tag{46}$$

Note that

$$f\left(y_i \mid x_i, \eta_i\right) = f\left(Hy_i \mid x_i, \eta_i\right)\left|\det\left(H\right)\right|, \tag{47}$$

but since $|\det(H)| = T^{-1/2}$ is a constant it can be ignored for our purposes. Moreover, since[4]

$$Cov\left(y_i^*, \overline{y}_i \mid x_i, \eta_i\right) = 0, \tag{48}$$

given normality we have that the conditional density of $y_i$ factorizes into the between-group and the orthogonal deviation densities:

$$f\left(y_i \mid x_i, \eta_i\right) = f\left(\overline{y}_i \mid x_i, \eta_i\right) f\left(y_i^* \mid x_i, \eta_i\right). \tag{49}$$

Note in addition that the orthogonal deviation density is independent of $\eta_i$

$$f\left(y_i^* \mid x_i, \eta_i\right) = f\left(y_i^* \mid x_i\right),$$

and in view of (40) it coincides with the conditional density given $\overline{y}_i$

$$f\left(y_i^* \mid x_i\right) = f\left(y_i \mid x_i, \overline{y}_i\right). \tag{50}$$

Thus, either way we have

$$\log f\left(y_i \mid x_i\right) = \log f\left(y_i^* \mid x_i\right) + \log \int f\left(\overline{y}_i \mid x_i, \eta_i\right) dF\left(\eta_i \mid x_i\right). \tag{51}$$

If $F\left(\eta_i \mid x_i\right)$ is unrestricted, the second term on the right-hand side of (51) is uninformative about $\beta$ so that the marginal ML estimators of $\beta$ and $\sigma^2$ coincide with the maximizers of $\sum_{i=1}^{N} \log f\left(y_i^* \mid x_i\right)$, which are again given by the WG estimator and (44). This is still true when $F\left(\eta_i \mid x_i\right)$ is specified to be Gaussian with unrestricted linear projection of $\eta_i$ on $x_i$, as in (45), but not when $\eta_i$ is assumed to be independent of $x_i$ (i.e. $\lambda = 0$), as we shall see in the next section.

---

[4] Note that

$$Cov\left(y_i^*, \overline{y}_i \mid x_i, \eta_i\right) = E\left(v_i^* \overline{v}_i \mid x_i, \eta_i\right) = AE\left(v_i v_i' \mid x_i, \eta_i\right) \iota / T = \sigma^2 A \iota / T = 0.$$

# 2 Error Components

The analysis in the previous section was motivated by the desire of identifying regression coefficients that are free from unobserved cross-sectional heterogeneity bias. Another major motivation for using panel data is the possibility of separating out permanent from transitory components of variation.

## 2.1 A Variance Decomposition

The starting point of our discussion is a simple variance-components model of the form

$$y_{it} = \mu + \eta_i + v_{it} \tag{52}$$

where $\mu$ is an intercept, $\eta_i \sim iid(0, \sigma_\eta^2)$, $v_{it} \sim iid(0, \sigma^2)$, and $\eta_i$ and $v_{it}$ are independent of each other. The cross-sectional variance of $y_{it}$ in any given period is given by $(\sigma_\eta^2 + \sigma^2)$. This model tells us that a fraction $\sigma_\eta^2/(\sigma_\eta^2 + \sigma^2)$ of the total variance corresponds to differences that remain constant over time while the rest are differences that vary randomly over time and units.

Dividing total variance into two components that are either completely fixed or completely random will often be unrealistic, but this model and its extensions are at the basis of much useful econometric descriptive work. A prominent example is the study of earnings inequality and mobility (cf. Lillard and Willis, 1978). In the analysis of transitions between log-normal earnings classes, the model allows us to distinguish between aggregate or unconditional transition probabilities and individual transition probabilities given certain values of permanent characteristics represented by $\eta_i$.

Indeed, given $\eta_i$, the $y$s are independent over time but with different means for different units, so that we have

$$y_i \mid \eta_i \sim id\left((\mu + \eta_i)\iota, \sigma^2 I_T\right),$$

whereas unconditionally we have

$$y_i \sim iid(\mu\iota, \Omega)$$

with

$$\Omega = \sigma^2 I_T + \sigma_\eta^2 \iota\iota'. \tag{53}$$

Thus the unconditional correlation between $y_{it}$ and $y_{is}$ for any two periods $t \neq s$ is given by

$$Corr(y_{it}, y_{is}) = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2} = \frac{\lambda}{1 + \lambda} \tag{54}$$

with $\lambda = \sigma_\eta^2/\sigma^2$.

**Estimating the Variance-Components Model** One possibility is to approach estimation conditionally given the $\eta_i$. That is, to estimate the realizations of the permanent effects that occur in the sample and $\sigma^2$. Natural unbiased estimates in this case would be

$$\widehat{\eta}_i = \overline{y}_i - \overline{y} \ (i = 1, ..., N) \tag{55}$$

and

$$\widehat{\sigma}^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \overline{y}_i)^2, \tag{56}$$

where $\overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ and $\overline{y} = N^{-1} \sum_{i=1}^{N} \overline{y}_i$. However, typically both $\sigma_\eta^2$ and $\sigma^2$ will be parameters of interest. To obtain an estimator of $\sigma_\eta^2$ note that the variance of $\overline{y}_i$ is given by

$$Var(\overline{y}_i) \equiv \overline{\sigma}^2 = \sigma_\eta^2 + \frac{\sigma^2}{T}. \tag{57}$$

Therefore, a large-$N$ consistent estimator of $\sigma_\eta^2$ can be obtained as the difference between the estimated variance of $\overline{y}_i$ and $\widehat{\sigma}^2/T$:

$$\widehat{\sigma}_\eta^2 = \frac{1}{N} \sum_{i=1}^{N} (\overline{y}_i - \overline{y})^2 - \frac{\widehat{\sigma}^2}{T}. \tag{58}$$

A problem with this estimator is that it is not guaranteed to be non-negative by construction.

The statistics (56) and (58) can be regarded as Gaussian ML estimates under $y_i \sim \mathcal{N}(\mu\iota, \Omega)$. To see this, note that using transformation (46) in general we have:

$$Hy_i = \begin{pmatrix} \overline{y}_i \\ y_i^* \end{pmatrix} \sim id \left[ \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \overline{\sigma}^2 & 0 \\ 0 & \sigma^2 I_{T-1} \end{pmatrix} \right]. \tag{59}$$

Hence, under normality the log density of $y_i$ can be decomposed as

$$\log f(y_i) = \log f(\overline{y}_i) + \log f(y_i^*), \tag{60}$$

so that the log likelihood of $(y_1, ..., y_N)$ is given by

$$L(\mu, \overline{\sigma}^2, \sigma^2) = L_B(\mu, \overline{\sigma}^2) + L_W(\sigma^2), \tag{61}$$

where

$$L_B(\mu, \overline{\sigma}^2) \propto -\frac{N}{2} \log \overline{\sigma}^2 - \frac{1}{2\overline{\sigma}^2} \sum_{i=1}^{N} (\overline{y}_i - \mu)^2 \tag{62}$$

and

$$L_W(\sigma^2) \propto -\frac{N(T-1)}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} y_i^{*\prime} y_i^*. \tag{63}$$

Clearly the ML estimates of $\sigma^2$ and $\overline{\sigma}^2$ are given by (56) and the sample variance of $\overline{y}_i$, respectively.[5] Moreover, the ML estimator of $\sigma_\eta^2$ is given by (58) in view of the invariance property of maximum likelihood estimation.

Note that with large $N$ and short $T$ we can obtain precise estimates of $\sigma_\eta^2$ and $\sigma^2$ but not of the individual realizations $\eta_i$. Conversely, with small $N$ and large $T$ we would be able to obtain accurate estimates of $\eta_i$ and $\sigma^2$ but not of $\sigma_\eta^2$, the intuition being that although we can estimate the individual $\eta_i$ well there may be too few of them to obtain a good estimate of their variance.

For large $N$, $\sigma_\eta^2$ is just-identified when $T = 2$ in which case we have $\sigma_\eta^2 = Cov(y_{i1}, y_{i2})$.[6]

## 2.2 Error-Components Regression

### 2.2.1 The Model

Often one is interested in the analysis of error-components models given some conditioning variables. The conditioning variables may be time-varying, time-invariant or both, denoted as $x_{it}$ and $f_i$, respectively. For example, we may be interested in separating out permanent and transitory components of individual earnings by labour market experience and educational categories.

This gives rise to a regression version of the previous model in which, in principle, not only $\mu$ but also $\sigma_\eta^2$ and $\sigma^2$ could be functions of $x_{it}$ and $f_i$. Nevertheless, in the standard error-components regression model $\mu$ is period-specific and made a linear function of $x_{it}$ and $f_i$, while the variance parameters are assumed not to vary with the regressors. The model is therefore

$$y_{it} = x_{it}'\beta + f_i'\gamma + u_{it} \tag{64}$$

$$u_{it} = \eta_i + v_{it} \tag{65}$$

together with the following assumption for the composite vector of errors $u_i = (u_{i1}, ..., u_{iT})'$:

$$u_i \mid w_i \sim iid(0, \sigma^2 I_T + \sigma_\eta^2 \iota\iota') \tag{66}$$

where $w_i = (x_{i1}', ..., x_{iT}', f_i')'$.

This model is similar to the one discussed in the previous chapter except in one fundamental aspect. The individual effect in the unobserved-heterogeneity model was potentially correlated with $x_{it}$. Indeed, this was the motivation for considering such a model in the first place. In contrast, in the error-components model $\eta_i$ and $v_{it}$ are two components of a regression error and hence both are uncorrelated with the regressors.

Formally, this model is a specialization of the unobserved-heterogeneity model of the previous chapter under Assumptions *A1* and *A2* in which in addition

$$E(\eta_i \mid w_i) = 0 \tag{67}$$

[5] Note that $\sum_{i=1}^N y_i^{*\prime} y_i^* = \sum_{i=1}^N \sum_{t=1}^T (y_{it} - \overline{y}_i)^2$.

[6] With $T = 2$, (58) coincides with the sample covariance between $y_{i1}$ and $y_{i2}$.

$$Var(\eta_i \mid w_i) = \sigma_\eta^2. \tag{68}$$

To reconcile the notation used in the two instances, note that in the unobserved heterogeneity model, the time-invariant component of the regression $f_i'\gamma$ is subsumed under the individual effect $\eta_i$. Moreover, in the unobserved-heterogeneity model we did not specify an intercept so that $E(\eta_i)$ was not restricted, whereas for the error-components model $E(\eta_i) = 0$, and $f_i$ will typically contain a constant term.

Note that in the error-components model $\beta$ is identified in a single cross-section. The parameters that require panel data for identification in this model are the variances of the components of the error $\sigma_\eta^2$ and $\sigma^2$, which typically will be parameters of central interest in this context.

There are also applications of model (64)-(65) in which the main interest lies in the estimation of $\beta$ and $\gamma$. In these cases it is natural to regard the error-components model as a restrictive version of the unobserved heterogeneity model of Section 1 with uncorrelated individual effects.

### 2.2.2 GLS and ML Estimation

Under the previous assumptions, OLS in levels provides unbiased and consistent but inefficient estimators of $\beta$ and $\gamma$:

$$\widehat{\delta}_{OLS} = \left(\sum_{i=1}^N W_i'W_i\right)^{-1} \sum_{i=1}^N W_i'y_i \tag{69}$$

where $W_i = \left(X_i \vdots \iota f_i'\right)$, $X_i = (x_{i1}, ..., x_{iT})'$, and $\delta = \left(\beta', \gamma'\right)'$.

Optimal estimation is achieved through GLS, also known as the Balestra–Nerlove estimator:[7]

$$\widehat{\delta}_{GLS} = \left(\sum_{i=1}^N W_i'\Omega^{-1}W_i\right)^{-1} \sum_{i=1}^N W_i'\Omega^{-1}y_i. \tag{70}$$

This GLS estimator is, nevertheless, unfeasible, since $\Omega$ depends on $\sigma_\eta^2$ and $\sigma^2$, which are unknown. Feasible GLS is obtained by replacing them by consistent estimates. Usually, the following are used:

$$\widehat{\sigma}^2 = \frac{1}{N(T-1)-k} \sum_{i=1}^N \sum_{t=1}^T \left(\widetilde{y}_{it} - \widetilde{x}_{it}'\widehat{\beta}_{WG}\right)^2 \tag{71}$$

$$\widehat{\sigma}_\eta^2 = \frac{1}{N} \sum_{i=1}^N \left(\overline{y}_i - \overline{w}_i'\widehat{\delta}_{BG}\right)^2 - \frac{\widehat{\sigma}^2}{T} \tag{72}$$

where $\widetilde{y}_{it} = y_{it} - \overline{y}_i$, $\widetilde{x}_{it} = x_{it} - \overline{x}_i$, and $\widehat{\delta}_{BG}$ denotes the *between-group* estimator, which is given by the OLS regression of $\overline{y}_i$ on $\overline{w}_i$:

$$\widehat{\delta}_{BG} = \left(\sum_{i=1}^N \overline{w}_i\overline{w}_i'\right) \sum_{i=1}^N \overline{w}_i\overline{y}_i. \tag{73}$$

[7] cf. Balestra and Nerlove (1966).

Alternatively, the full set of parameters $\beta$, $\gamma$, $\sigma^2$, and $\sigma^2_\eta$ may be jointly estimated by maximum likelihood. As in the case without regressors, the log likelihood can be decomposed as the sum of the *between* and *within* log likelihoods. In this case we have:

$$\begin{pmatrix} \overline{y}_i \\ y_i^* \end{pmatrix} \mid w_i \sim id \left[ \begin{pmatrix} \overline{x}_i'\beta + f_i'\gamma \\ X_i^*\beta \end{pmatrix}, \begin{pmatrix} \overline{\sigma}^2 & 0 \\ 0 & \sigma^2 I_{T-1} \end{pmatrix} \right], \tag{74}$$

so that under normality the error-components log likelihood equals:

$$L\left(\beta, \gamma, \sigma^2, \overline{\sigma}^2\right) = L_B\left(\beta, \gamma, \overline{\sigma}^2\right) + L_W\left(\beta, \sigma^2\right) \tag{75}$$

where

$$L_B\left(\beta, \gamma, \overline{\sigma}^2\right) \propto -\frac{N}{2}\log \overline{\sigma}^2 - \frac{1}{2\overline{\sigma}^2}\sum_{i=1}^{N}\left(\overline{y}_i - \overline{x}_i'\beta - f_i'\gamma\right)^2 \tag{76}$$

and

$$L_W\left(\beta, \sigma^2\right) \propto -\frac{N(T-1)}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i^* - X_i^*\beta\right)'\left(y_i^* - X_i^*\beta\right). \tag{77}$$

Separate maximization of $L_W$ and $L_B$ gives rise to within-group and between group estimation, respectively. Thus, the error-components likelihood can be regarded as enforcing the restriction that the parameter vectors $\beta$ that appear in $L_W$ and $L_B$ coincide. This immediately suggests a (likelihood-ratio) specification test that will be further discussed below.

Moreover, in the absence of individual effects $\sigma^2_\eta = 0$ so that $\overline{\sigma}^2 = \sigma^2/T$. Thus, the OLS estimator in levels (69) can be regarded as the MLE that maximizes the log-likelihood (75) subject to the restriction $\overline{\sigma}^2 = \sigma^2/T$. Again, this suggests a likelihood-ratio (LR) test of the existence of (uncorrelated) effects based on the comparison of the restricted and unrestricted likelihoods. Such a test will, nevertheless, be sensitive to distributional assumptions.

In terms of the transformed model, $\widehat{\delta}_{GLS}$ can be written as a weighted least-squares estimator:

$$\widehat{\delta}_{GLS} = \left[\sum_{i=1}^{N}\left(W_i^{*\prime}W_i^* + \phi^2\overline{w}_i\overline{w}_i'\right)\right]^{-1}\sum_{i=1}^{N}\left(W_i^{*\prime}y_i^* + \phi^2\overline{w}_i\overline{y}_i\right) \tag{78}$$

where $\phi^2$ is the ratio of the within to the between error variances $\phi^2 = \sigma^2/\overline{\sigma}^2$, $W_i^* = AW_i$, and $\overline{w}_i = T^{-1}W_i'\iota$. Thus $\widehat{\delta}_{GLS}$ can be regarded as a matrix-weighted average of the within-group and between-group estimators (Maddala, 1971). The statistic (78) is identical to (70).[8] For feasible GLS, $\phi^2$ is replaced by the ratio of the within to the between sample residual variances $\widehat{\phi}^2 = \widehat{\sigma}^2/\widehat{\overline{\sigma}}^2$.

So far we have motivated error-components regression models from a direct interest in the components themselves. Sometimes, however, correlation between individual effects and regressors can be regarded as an empirical issue. Next we address the testing of such hypothesis.

[8] When $\phi^2 = T$ (or $\sigma^2_\eta = 0$) (78) boils down to the OLS in levels estimator (69), whereas if $\sigma^2_\eta \to \infty$ then $\phi^2 \to 0$ and $\widehat{\delta}_{GLS}$ tends to within-groups.

## 2.3 Testing for Correlated Unobserved Heterogeneity

Sometimes correlated unobserved heterogeneity is a basic property of the model of interest. An example is a "$\lambda$-constant" labour supply equation where $\eta_i$ is determined by the marginal utility of initial wealth, which according to the underlying life-cycle model will depend on wages in all periods (MaCurdy, 1981). Another example is when a regressor is a lagged dependent variable. In cases like this, testing for lack of correlation between regressors and individual effects is not warranted since we wish the model to have this property.

On other occasions, correlation between regressors and individual effects can be regarded as an empirical issue. In these cases testing for correlated unobserved heterogeneity can be a useful *specification test* for regression models estimated in levels. Researchers may have a preference for models in levels because estimates in levels are in general more precise than estimates in deviations (dramatically so when the time series variation in the regressors relative to the cross-sectional variation is small), or because of an interest in regressors that lack time series variation.

### 2.3.1 Specification Tests

We have already suggested a specification test of correlated effects from a likelihood ratio perspective. This was a test of equality of the $\beta$ coefficients appearing in the WG and BG likelihoods. Similarly, from a least-squares perspective, we may consider the system

$$\overline{y}_i = \overline{x}_i' b + f_i' c + \varepsilon_i \tag{79}$$

$$y_i^* = X_i^* \beta + u_i^*, \tag{80}$$

where $b$, $c$, and $\varepsilon_i$ are such that $E^*(\varepsilon_i \mid \overline{x}_i, f_i) = 0$, and formulate the problem as a (Wald) test of the null hypothesis[9]

$$H_0: \ \beta = b. \tag{81}$$

The least-squares perspective is of interest because it can easily accommodate robust generalizations to heteroskedasticity and serial correlation.

Under the unobserved-heterogeneity model

$$E(\overline{y}_i \mid w_i) = \overline{x}_i' \beta + f_i' \gamma + E(\eta_i \mid w_i),$$

so that (79) can be regarded as a specification of an alternative hypothesis of the form

$$H_1: \ E(\eta_i \mid w_i) = \overline{x}_i' \lambda_1 + f_i' \lambda_2 \tag{82}$$

with $b = \beta + \lambda_1$ and $c = \gamma + \lambda_2$. $H_0$ is, therefore, equivalent to $\lambda_1 = 0$. Note that $H_0$ does not specify that $\lambda_2 = 0$, which is not testable.

[9] Under the assumptions of the error-components model $b = \beta$, $c = \gamma$, and $\varepsilon_i = \overline{u}_i$.

Under (82) and the additional assumption $Var(\eta_i \mid w_i) = \sigma_\eta^2$, the error covariance matrix of the system (79)-(80) is given by $Var(\varepsilon_i \mid w_i) = \overline{\sigma}^2$, $Cov(\varepsilon_i, u_i^* \mid w_i) = 0$, and $Var(u_i^* \mid w_i) = \sigma^2 I_{T-1}$. Thus the optimal LS estimates of $(b', c')'$ and $\beta$ are the BG and the WG estimators, respectively. Explicit expressions for the BG estimator of $b$ and its estimated variance matrix are:

$$\widehat{b}_{BG} = \left(\overline{X}'M\overline{X}\right)^{-1}\overline{X}'M\overline{y} \tag{83}$$

$$\widehat{V}_{BG} \equiv \widehat{Var}\left(\widehat{b}_{BG}\right) = \widehat{\overline{\sigma}}^2\left(\overline{X}'M\overline{X}\right)^{-1} \tag{84}$$

where $M = I - F\left(F'F\right)^{-1}F'$, $F = (f_1, ..., f_N)'$, $\overline{X} = (\overline{x}_1, ..., \overline{x}_N)'$, and $\overline{y} = (\overline{y}_1, ..., \overline{y}_N)'$. Likewise, the estimated variance matrix of the WG estimator is

$$\widehat{V}_{WG} \equiv \widehat{Var}\left(\widehat{\beta}_{WG}\right) = \widehat{\sigma}^2\left(\sum_{i=1}^N X_i^{*\prime}X_i^*\right)^{-1}. \tag{85}$$

Moreover, since $Cov\left(\widehat{b}_{BG}, \widehat{\beta}_{WG}\right) = 0$, the Wald test of (81) is given by

$$h = \left(\widehat{b}_{BG} - \widehat{\beta}_{WG}\right)'(\widehat{V}_{WG} + \widehat{V}_{BG})^{-1}\left(\widehat{b}_{BG} - \widehat{\beta}_{WG}\right). \tag{86}$$

Under $H_0$, the statistic $h$ will have a $\chi^2$ distribution with $k$ degrees of freedom in large samples. Clearly, $h$ will be sensitive to the nature of the variables included in $f_i$. For example, $H_0$ might be rejected when $f_i$ only contains a constant term, but not when a larger set of time-invariant regressors is included.

Hausman (1978) originally motivated the testing of correlated effects as a comparison between WG and the Balestra–Nerlove GLS estimator, suggesting a statistic of the form

$$h = \left(\widehat{\beta}_{GLS} - \widehat{\beta}_{WG}\right)'(\widehat{V}_{WG} - \widehat{V}_{GLS})^{-1}\left(\widehat{\beta}_{GLS} - \widehat{\beta}_{WG}\right), \tag{87}$$

where

$$\widehat{V}_{GLS} = \widehat{\sigma}^2(X^{*\prime}X^* + \widehat{\phi}^2\overline{X}'M\overline{X})^{-1}. \tag{88}$$

Under $H_0$ both estimators are consistent, so we would expect the difference $\widehat{\beta}_{GLS} - \widehat{\beta}_{WG}$ to be small. Moreover, since $\widehat{\beta}_{GLS}$ is efficient, the variance of the difference must be given by the difference of variances. Otherwise, we could find a linear combination of the two estimators that would be more efficient than GLS. Under $H_1$ the WG estimator remains consistent but GLS does not, so their difference and the test statistic will tend to be large. A statistic of the form given in (87) is known as a Hausman test statistic. As shown by Hausman and Taylor (1981), (87) is in fact the same statistic as (86). Thus $h$ can be regarded both as a Hausman test or as a Wald test of the restriction $\lambda_1 = 0$ from OLS estimates of the model under the alternative.

If the errors are heteroskedastic and/or serially correlated, the previous formulae for the large sample variances of WG, BG, and GLS are not valid. Moreover, WG and GLS cannot be ranked in terms of efficiency so that the variance of the difference between the two does not coincide with the difference of variances. Following the Wald approach, Arellano (1993) suggested a generalized test that is robust to heteroskedasticity and autocorrelation.

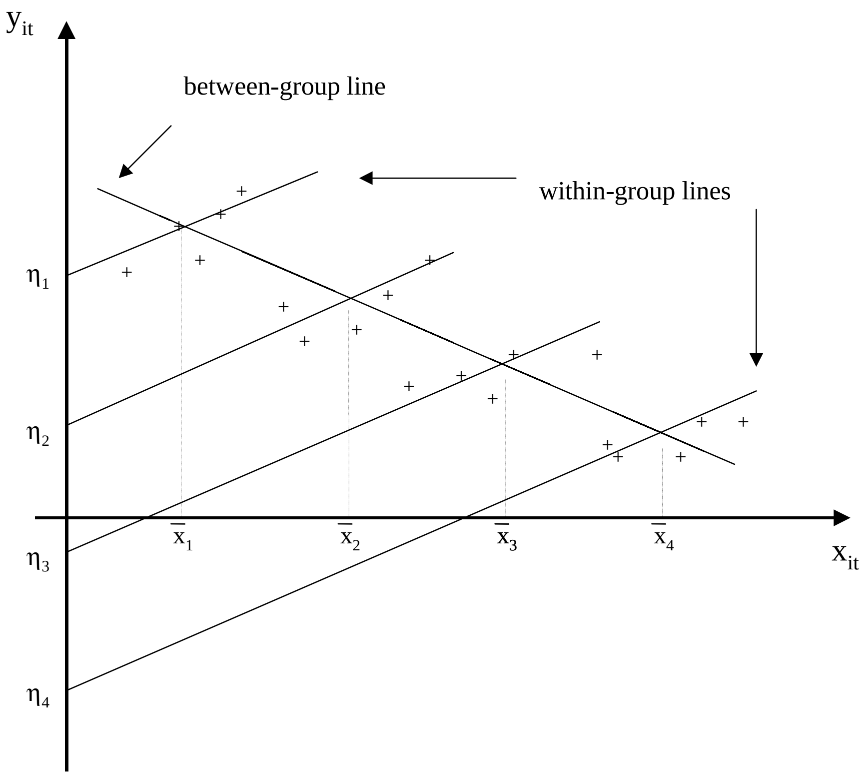


Figure 1: Within-group and between-group lines

**Fixed Effects *versus* Random Effects** These specification tests are sometimes described as tests of random effects against fixed effects. However, according to the previous discussion, for typical econometric panels, we shall not be testing the nature of the sampling process but the dependence between individual effects and regressors. Thus, for our purposes individual effects may be regarded as random without loss of generality. Provided the interest is in partial regression coefficients holding effects constant, what matters is whether the effects are independent of observed regressors or not.

Figure 1 provides a simple illustration for the scatter diagram of a panel data set with $N = 4$ and $T = 5$. In this example there is a marked difference between the positive slope of the within-group lines and the negative one of the between-group regression. This situation is the result of the strong negative association between the individual intercepts and the individual averages of the regressors.

### 2.3.2 Robust GMM Estimation and Testing

Under the null of uncorrelated effects we may consider GMM estimation based on the orthogonality conditions[10]

$$E\left[x_i\left(\overline{y}_i - \overline{x}_i'\beta - f_i'\gamma\right)\right] = 0 \tag{89}$$

$$E\left[f_i\left(\overline{y}_i - \overline{x}_i'\beta - f_i'\gamma\right)\right] = 0 \tag{90}$$

$$E\left[(y_i^* - X_i^*\beta) \otimes x_i\right] = 0. \tag{91}$$

In parallel with the development in Section 1.3.3, the resulting estimates of $\beta$ and $\gamma$ will be asymptotically equivalent to Balestra–Nerlove GLS with classical errors, but strictly more efficient when heteroskedasticity or autocorrelation is present. However, under the alternative of correlated effects, any GMM estimate that relies on the moments (89) will be inconsistent for $\beta$. Thus, we may test for correlated effects by considering an incremental test of the over identifying restrictions (89). Note that under the alternative, GMM estimates based on (90)-(91) will be consistent for $\beta$ but not necessarily for $\gamma$.

Optimal GMM estimates in this context minimize a criterion of the form

$$s(\delta) = \left[\sum_{i=1}^{N}(y_i - W_i\delta)H'Z_i\right]\left(\sum_{i=1}^{N}Z_i'H\widehat{u}_i\widehat{u}_i'H'Z_i\right)^{-1}\left[\sum_{i=1}^{N}Z_i'H(y_i - W_i\delta)\right] \tag{92}$$

where $H\widehat{u}_i$ are some one-step consistent residuals. Under uncorrelated effects the instrument matrix $Z_i$ takes the form

$$Z_i = \begin{pmatrix} x_i' & f_i' & 0 \\ 0 & 0 & I_{T-1} \otimes x_i' \end{pmatrix}, \tag{93}$$

whereas under correlated effects we shall use

$$Z_i = \begin{pmatrix} f_i' & 0 \\ 0 & I_{T-1} \otimes x_i' \end{pmatrix}. \tag{94}$$

[10] We could also add:

$$E\left[(y_i^* - X_i^*\beta) \otimes f_i\right] = 0,$$

in which case, the entire set of moments can be expressed in terms of the original equation system as:

$$E\left[\left(y_i - X_i\beta - \iota f_i'\gamma\right) \otimes w_i\right] = 0.$$

When $f_i$ contains a constant term only, this amounts to including a set of time dummies in the instrument set.

## 2.4 Models with Information in Levels

Sometimes it is of central interest to measure the effect of a time-invariant explanatory variable controlling for unobserved heterogeneity. Returns to schooling holding unobserved ability constant is a prominent example. In those cases, as explained in Section 1, panel data is not directly useful. Hausman and Taylor (1981) argued, however, that panel data might still be useful in an indirect way if the model contained time-varying explanatory variables that were uncorrelated with the effects.

Suppose there are subsets of the time-invariant and time-varying explanatory variables, $f_{1i}$ and $x_{1i} = (x'_{1i1}, ..., x'_{1iT})'$ respectively, that can be assumed a priori to be uncorrelated with the effects. In such case, the following subset of the orthogonality conditions (89)-(91) hold

$$E\left[x_{1i}\left(\overline{y}_i - \overline{x}'_i\beta - f'_i\gamma\right)\right] = 0 \tag{95}$$

$$E\left[f_{1i}\left(\overline{y}_i - \overline{x}'_i\beta - f'_i\gamma\right)\right] = 0 \tag{96}$$

$$E\left[(y^*_i - X^*_i\beta) \otimes x_i\right] = 0. \tag{97}$$

The parameter vector $\beta$ will be identified from the moments for the errors in deviations (97). The basic point noted by Hausman and Taylor is that the coefficients $\gamma$ may also be identified using the variables $x_{1i}$ and $f_{1i}$ as instruments for the errors in levels, provided the rank condition is satisfied. Given identification, the coefficients $\beta$ and $\gamma$ can be estimated by GMM (Arellano and Bover, 1995).

The notion that a time-varying variable that is uncorrelated with an individual effect can be used at the same time as an instrument for itself and for a correlated time-invariant variable is potentially appealing. Nevertheless, the impact of these models in applied work has been limited, due to the difficulty in finding exogenous variables that can be convincingly regarded a priori as being uncorrelated with the individual effects.

# 3 Error in Variables

## 3.1 Introduction to the Standard Regression Model with Errors in Variables

Let us consider a cross-sectional regression model

$$y_i = \alpha + x_i^{\dagger}\beta + v_i. \tag{98}$$

Suppose we actually observe $y_i$ and $x_i$, which is a noisy measure of $x_i^{\dagger}$ subject to an additive measurement error $\varepsilon_i$

$$x_i = x_i^{\dagger} + \varepsilon_i. \tag{99}$$

We assume that all the unobservables $x_i^{\dagger}$, $v_i$, and $\varepsilon_i$ are mutually independent with variances $\sigma_{\dagger}^2$, $\sigma_v^2$, and $\sigma_{\varepsilon}^2$. Since $v_i$ is independent of $x_i^{\dagger}$, $\beta$ is given by the population regression coefficient of $y_i$ on $x_i^{\dagger}$:

$$\beta = \frac{Cov(y_i, x_i^{\dagger})}{Var(x_i^{\dagger})}, \tag{100}$$

but since $x_i^{\dagger}$ is unobservable we cannot use a sample counterpart of this expression as an estimator of $\beta$.

What do we obtain by regressing $y_i$ on $x_i$ in the population? The result is

$$\frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(y_i, x_i^{\dagger} + \varepsilon_i)}{\sigma_{\dagger}^2 + \sigma_{\varepsilon}^2} = \frac{Cov(y_i, x_i^{\dagger})}{\sigma_{\dagger}^2 + \sigma_{\varepsilon}^2} = \frac{\beta}{1+\lambda} \tag{101}$$

where $\lambda = \sigma_{\varepsilon}^2/\sigma_{\dagger}^2$. Note that since $\lambda$ is non-negative by construction, the population regression coefficient of $y_i$ on $x_i$ will always be smaller than $\beta$ in absolute value as long as $\sigma_{\varepsilon}^2 > 0$.

This type of model is relevant in at least two conceptually different situations. One corresponds to instances of actual measurement errors due to misreporting, rounding-off errors, etc. The other arises when the variable of economic interest is a latent variable which does not correspond exactly to the one that is available in the data.

In either case, the worry is that the variable to which agents respond does not coincide with the one that is entered as a regressor in the model. The result is that the unobservable component in the relationship between $y_i$ and $x_i$ will contain a multiple of the measurement error in addition to the error term in the original relationship:

$$y_i = \alpha + x_i\beta + u_i \tag{102}$$

$$u_i = v_i - \beta\varepsilon_i. \tag{103}$$

Clearly, the observed regressor $x_i$ will be correlated with $u_i$ even if the latent variable $x_i^{\dagger}$ is not.

This problem is often of practical significance, specially in regression analysis with micro data, since the resulting biases may be large. Note that the magnitude of the bias does not depend on the absolute magnitude of the measurement error variance but on the "noise to signal" ratio $\lambda$. For example, if $\lambda = 1$, so that 50 per cent of the total variance observed in $x_i$ is due to measurement error—which is not an uncommon situation—the population regression coefficient of $y_i$ on $x_i$ will be half the value of $\beta$.

As for solutions, suppose we have the means of assessing the extent of the measurement error, so that $\lambda$ or $\sigma_\varepsilon^2$ are known or can be estimated. Then $\beta$ can be determined as

$$\beta = (1+\lambda)\frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov(y_i, x_i)}{Var(x_i) - \sigma_\varepsilon^2}. \tag{104}$$

More generally, in a model with $k$ regressors and a conformable vector of coefficients $\beta$

$$y_i = x_i'\beta + \left(v_i - \varepsilon_i'\beta\right) \tag{105}$$

with $E(\varepsilon_i \varepsilon_i') = \Omega_\varepsilon$, $E(x_i^\dagger x_i^{\dagger\prime}) = \Omega_\dagger$ and $\Lambda = \Omega_\dagger^{-1}\Omega_\varepsilon$:

$$\beta = (I_k + \Lambda)\left[E(x_i x_i')\right]^{-1} E(x_i y_i) = \left[E(x_i x_i') - \Omega_\varepsilon\right]^{-1} E(x_i y_i). \tag{106}$$

In this notation, $x_i$ will include a constant term, and possibly other regressors without measurement error. This situation will be reflected by the occurrence of zeros in the corresponding elements of $\Omega_\varepsilon$.

The expression (106) suggests an estimator of the form

$$\widetilde{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i' - \widetilde{\Omega}_\varepsilon^{-1}\right)\frac{1}{N}\sum_{i=1}^{N} x_i y_i. \tag{107}$$

where $\widetilde{\Omega}_\varepsilon$ denotes a consistent estimate of $\Omega_\varepsilon$.

Alternatively, if we have a second noisy measure of $x_i^\dagger$

$$z_i = x_i^\dagger + \zeta_i \tag{108}$$

such that the measurement error $\zeta_i$ is independent of $\varepsilon_i$ and the other unobservables, it can be used as an instrumental variable. In effect, for scalar $x_i$ we have

$$\frac{Cov(z_i, y_i)}{Cov(z_i, x_i)} = \frac{Cov(x_i^\dagger + \zeta_i, y_i)}{Cov(x_i^\dagger + \zeta_i, x_i^\dagger + \varepsilon_i)} = \frac{Cov(y_i, x_i^\dagger)}{Var(x_i^\dagger)} = \beta. \tag{109}$$

Moreover, since also

$$\frac{Cov(x_i, y_i)}{Cov(x_i, z_i)} = \beta, \tag{110}$$

there is one overidentifying restriction in this problem.

In some way the instrumental variable solution is not different from the previous one. Indirectly, the availability of two noisy measures is used to identify the systematic and measurement error variances. Note that since

$$Var\begin{pmatrix} x_i \\ z_i \end{pmatrix} = \begin{pmatrix} \sigma_\dagger^2 + \sigma_\varepsilon^2 & \sigma_\dagger^2 \\ \sigma_\dagger^2 & \sigma_\dagger^2 + \sigma_\zeta^2 \end{pmatrix} \tag{111}$$

we can determine the variances of the unobservables as

$$\sigma_\dagger^2 = Cov(z_i, x_i) \tag{112}$$

$$\sigma_\varepsilon^2 = Var(x_i) - Cov(z_i, x_i) \tag{113}$$

$$\sigma_\zeta^2 = Var(z_i) - Cov(z_i, x_i). \tag{114}$$

In econometrics the instrumental variable approach is the most widely used technique. Thus, the response to measurement error bias in linear regression problems is akin to the response to simultaneity bias. This similarity, however, no longer holds in the nonlinear regression context. The problem is that in a nonlinear regression the measurement error is no longer additively separable from the true value of the regressor.

### 3.2 Measurement Error Bias and Unobserved Heterogeneity Bias

Let us consider a cross-sectional model that combines measurement error and unobserved heterogeneity

$$y_i = x_i^\dagger \beta + \eta_i + v_i \tag{115}$$

$$x_i = x_i^\dagger + \varepsilon_i,$$

where all unobservables are independent, except $x_i^\dagger$ and $\eta_i$. The population regression coefficient of $y_i$ on $x_i$ is given by

$$\frac{Cov(y_i, x_i)}{Var(x_i)} = \beta + \frac{Cov(\eta_i + v_i - \beta\varepsilon_i, x_i)}{Var(x_i)} = \beta - \left(\frac{\sigma_\varepsilon^2}{\sigma_\dagger^2 + \sigma_\varepsilon^2}\right)\beta + \left(\frac{Cov(\eta_i, x_i^\dagger)}{\sigma_\dagger^2 + \sigma_\varepsilon^2}\right). \tag{116}$$

Note that there are two components to the bias. The first one is due to measurement error and depends on $\sigma_\varepsilon^2$. The second is due to unobserved heterogeneity and depends on $Cov(\eta_i, x_i^\dagger)$. Sometimes these two biases tend to offset each other. For example, if $\beta > 0$ and $Cov(\eta_i, x_i^\dagger) > 0$, the measurement error bias will be negative while the unobserved heterogeneity bias will be positive. A full offsetting would only occur if $Cov(\eta_i, x_i^\dagger) = \sigma_\varepsilon^2 \beta$, something that could only happen by chance.

**Measurement Error Bias in First-Differences** Suppose we have panel data with $T = 2$ and consider a regression in first-differences as a way of removing unobserved heterogeneity bias. In such a case we obtain

$$\frac{Cov(\Delta y_{i2}, \Delta x_{i2})}{Var(\Delta x_{i2})} = \frac{\beta}{1 + \lambda_\Delta} \tag{117}$$

where $\lambda_\Delta = Var(\Delta\varepsilon_{i2})/Var(\Delta x_{i2}^\dagger)$.

The main point to make here is that first-differencing may exacerbate the measurement error bias. The reason is as follows. If $\varepsilon_{it}$ is an *iid* error then $Var(\Delta\varepsilon_{i2}) = 2\sigma_\varepsilon^2$. If $x_{it}^\dagger$ is also *iid* then $\lambda_\Delta = \lambda$, and the measurement error bias in levels and first-differences will be of the same magnitude. However, if $x_{it}^\dagger$ is a stationary time series with positive serial correlation

$$Var(\Delta x_{i2}^\dagger) = 2\left[\sigma_\dagger^2 - Cov(x_{i1}^\dagger, x_{i2}^\dagger)\right] < 2\sigma_\dagger^2 \tag{118}$$

and therefore $\lambda_\Delta > \lambda$.[11]

A related example of this situation in data with a group structure arises in the analysis of the returns to schooling with data on twin siblings. Regressions in differences remove genetic ability bias but may exacerbate measurement error bias in schooling if the siblings' measurement errors are independent but their true schooling attainments are highly correlated (Griliches, 1977).

Under the same circumstances, the within-group measurement error bias with $T > 2$ will be smaller than that in first-differences but higher than the measurement error bias in levels (Griliches and Hausman, 1986).

Therefore, the finding of significantly different results in regressions in first-differences and orthogonal deviations may be an indication of the presence of measurement error.

### 3.3 Instrumental Variable Estimation with Panel Data

The availability of panel data helps to solve the problem of measurement error bias by providing internal instruments as long as we are willing to restrict the serial dependence in the measurement error.

In a model without unobserved heterogeneity the following orthogonality conditions are valid pro-

[11] Note that, as explained in Section 2, the cross-sectional covariance between $x_{i1}^\dagger$ and $x_{i2}^\dagger$ will also be positive in the presence of heterogeneity, even if the individual time series are not serially correlated.

vided the measurement error is white noise and $T \geq 2$:

$$E\left[\begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{i(t-1)} \\ x_{i(t+1)} \\ \vdots \\ x_{iT} \end{pmatrix} (y_{it} - \alpha - x_{it}\beta)\right] = 0 \;(t = 1, ..., T). \tag{119}$$

Note that this situation is compatible with the presence of serial correlation in the disturbance term in the relationship between $y$ and $x$. This is so because the disturbance is made of two components:

$$u_{it} = v_{it} - \varepsilon_{it}\beta$$

and only the second is required to be white noise for the validity of the moment conditions above.

Also note that identification of $\beta$ from the previous moments requires that $x_{it}$ is predictable from its past and future values. Thus, the rank condition for identification would fail if the latent variable $x_{it}^{\dagger}$ was also white noise.

In a model with unobserved heterogeneity and a white noise measurement error, we can rely on the following moments for the errors in first-differences provided $T \geq 3$:[12]

$$E\left[\begin{pmatrix} x_{i1} \\ \vdots \\ x_{i(t-2)} \\ x_{i(t+1)} \\ \vdots \\ x_{iT} \end{pmatrix} (\Delta y_{it} - \Delta x_{it}\beta)\right] = 0 \;(t = 2, ..., T). \tag{120}$$

Moments of this type and GMM estimators based on them were proposed by Griliches and Hausman (1986).

With $T = 3$ we would have the following two orthogonality conditions:

$$E\left[x_{i3}\left(\Delta y_{i2} - \Delta x_{i2}\beta\right)\right] = 0 \tag{121}$$

$$E\left[x_{i1}\left(\Delta y_{i3} - \Delta x_{i3}\beta\right)\right] = 0. \tag{122}$$

As in the previous case, if $x_{it}^{\dagger}$ were white noise the rank condition for identification would not be satisfied. Also, if $x_{it}^{\dagger}$ was a random walk then $Cov(x_{i1}, \Delta x_{i3}) = 0$ but $Cov(x_{i3}, \Delta x_{i2}) \neq 0$. Note

[12] In this discussion we use first-differences to remove individual effects. Note that the use of forward orthogonal deviations would preclude the use of future values of $x$ as instruments.

that these instrumental variable methods can be expected to be useful in the same circumstances under which differencing exacerbates measurement error bias. Namely, when there is more time series dependence in $x^{\dagger}_{it}$ than in $\varepsilon_{it}$.

If measured persistence in $x^{\dagger}_{it}$ is exclusively due to unobserved heterogeneity, however, the situation is not different from the homogeneous white noise case and the rank condition will still fail. Specifically, suppose that

$$x^{\dagger}_{it} = \mu_i + \xi_{it} \tag{123}$$

where $\xi_{it}$ is *iid* over $i$ and $t$, and independent of $\mu_i$. Then $Cov(x_{i1}, \Delta x_{i3}) = Cov(x_{i3}, \Delta x_{i2}) = 0$, with the result that $\beta$ is unidentifiable from (121) and (122). This situation was discussed by Chamberlain (1984, 1985) who noted the observational equivalence between the measurement error and the fixed effects models when the process for $x^{\dagger}_{it}$ is as in (123).

Finally, note that the assumptions about the measurement error properties can be relaxed somewhat provided the panel is sufficiently long and there is suitable dependence in the latent regressor. For example, $\varepsilon_{it}$ could be allowed to be a first-order moving average process in which case the valid instruments in the first-difference equation for period $t$ would be

$$\left(x_{i1}, ..., x_{i(t-3)}, x_{i(t+2)}, ..., x_{iT}\right). \tag{124}$$

### 3.4 Illustration: Measuring Economies of Scale in Firm Money Demand

As an illustration of the previous discussion, we report some estimates from Bover and Watson (2005) concerning economies of scale in a firm money demand equation of the type discussed in Section 1.

The equations estimated by Bover and Watson are of the general form given in (5):

$$\log m_{it} = c(t) \log s_{it} + b(t) + \eta_i + v_{it}. \tag{125}$$

The scale coefficient $c(t)$ is specified as a second-order polynomial in $t$ to allow for changes in economies of scale over the sample period. The year dummies $b(t)$ capture changes in relative interest rates together with other aggregate effects. The individual effect is meant to represent permanent differences across firms in the production of transaction services (so that $\eta = -\log a$), and $v$ contains measurement errors in cash holdings and sales. We would expect a non-negative correlation between sales and $a$, implying $Cov\left(\log s, \eta\right) \leq 0$ and a downward unobserved heterogeneity bias in economies of scale.

Table 1
Firm Money Demand Estimates
Sample period 1986–1996

| | OLS Levels | OLS Orthogonal deviations | OLS 1st-diff. | GMM 1st-diff. | GMM 1st-diff. m. error | GMM Levels m. error |
|---|---|---|---|---|---|---|
| Log sales | .72 | .56 | .45 | .49 | .99 | .75 |
| | (30.) | (16.) | (12.) | (16.) | (7.5) | (35.) |
| Log sales ×trend | −.02 | −.03 | −.03 | −.03 | −.03 | −.03 |
| | (3.2) | (9.7) | (4.9) | (5.3) | (5.0) | (4.0) |
| Log sales ×trend$^2$ | .001 | .002 | .001 | .001 | .001 | .001 |
| | (1.2) | (6.6) | (1.9) | (2.0) | (2.3) | (1.4) |
| Sargan ($p$-value) | | | | .12 | .39 | .00 |

All estimates include year dummies, and those in levels also include industry dummies. $t$-ratios in brackets robust to heteroskedasticity & serial correlation. $N$=5649. Source: Bover and Watson (2005).

All the estimates in Table 1 are obtained from an unbalanced panel of 5649 Spanish firms with at least four consecutive annual observations during the period 1986−1996.[13]

The comparison between OLS in levels and orthogonal deviations (columns 1 and 2) is consistent with a positive unobserved heterogeneity bias (the opposite to what we expected), but the smaller sales effect obtained by OLS in first-differences (column 3) suggests that measurement error bias may be important.

Column 4 shows two-step robust GMM estimates based on the moments $E\left(\log s_{it}\Delta v_{is}\right) = 0$ for all $t$ and $s$ (in addition to time dummies). These estimates are of the form given in (34) with weight matrix (35). In the absence of measurement error, we would expect them to be consistent for the same parameters as OLS in orthogonal deviations and first-differences. In fact, in the case of Table 1 the last two differ, the GMM sales coefficient lies between the two, and the test statistic of overidentifying restrictions (Sargan) is marginal.

[13] The use of an unbalanced panel requires the introduction of some modifications in the formulae for the estimators, which we do not consider here.

Column 5 shows GMM estimates based on

$$E\left(\log s_{it}\Delta v_{is}\right)=0\ (t=1,...,s-2,s+1,..,T;s=1,...,T), \tag{126}$$

thus allowing for both correlated firm effects and serially independent multiplicative measurement errors in sales. Interestingly, now the leading sales coefficient is much higher and close to unity, and the Sargan test has a $p$-value close to 40 per cent.

Finally, column 6 shows GMM estimates based on

$$E\left(\log s_{it}v_{is}\right)=0\ (t=1,...,s-1,s+1,..,T;s=1,...,T). \tag{127}$$

In this case, as with the other estimates in levels, firm effects in (125) are replaced by industry effects. Therefore, the estimates in column 6 allow for serially uncorrelated measurement error in sales but not for correlated effects. The leading sales effect in this case is close to OLS in levels, suggesting that in levels the measurement error bias is not as important as in the estimation in differences. The Sargan test provides a sound rejection, which can be interpreted as a rejection of the null of lack of correlation between sales and firm effects, allowing for measurement error.

What is interesting about this example is that a comparison between estimates in levels and deviations without consideration of the possibility of measurement error (e.g. restricted to compare columns 1 and 2, or 1 and 3, as in Hausman-type testing), would lead to the conclusion of correlated effects, but with biases going in entirely the wrong direction.

# References


[1] Arellano, M. (1987): "Computing Robust Standard Errors for Within-Group Estimators", *Oxford Bulletin of Economics and Statistics*, 49, 431-434.

[2] Arellano, M. (1993): "On the Testing of Correlated Effects with Panel Data", *Journal of Econometrics*, 59, 87-97.

[3] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press, Oxford.

[4] Arellano, M. and O. Bover (1995): "Another Look at the Instrumental-Variable Estimation of Error-Components Models", *Journal of Econometrics*, 68, 29-51.

[5] Balestra, P. and M. Nerlove (1966): "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas", *Econometrica*, 34, 585-612.

[6] Bover, O. and N. Watson (2005): "Are There Economies of Scale in the Demand for Money by Firms? Some Panel Data Estimates", *Journal of Monetary Economics*, 52, 1569-1589.

[7] Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data", *Journal of Econometrics*, 18, 5-46.

[8] Chamberlain, G. (1984): "Panel Data", in Griliches, Z. and M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

[9] Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence", in Heckman, J. J. and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge.

[10] Griliches, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, 45, 1-22.

[11] Griliches, Z. and J. A. Hausman (1986): "Errors in Variables in Panel Data", *Journal of Econometrics*, 31, 93-118.

[12] Hausman, J. A. (1978): "Specification Tests in Econometrics", *Econometrica*, 46, 1251-1272.

[13] Hausman, J. A. and W. E. Taylor (1981): "Panel Data an Unobservable Individual Effects", *Econometrica*, 49, 1377-1398.

[14] Hoch, I. (1962): "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data", *Econometrica*, 30, 34-53.

[15] Kiefer, N. M. (1980): "Estimation of Fixed Effect Models for Time Series of Cross-Sections with Arbitrary Intertemporal Covariance", *Journal of Econometrics*, 14, 195-202.

[16] Lillard, L. and R. J. Willis (1978): "Dynamic Aspects of Earnings Mobility", *Econometrica*, 46, 985-1012.

[17] MaCurdy, T. E. (1981): "An Empirical Model of Labor Supply in a Life-Cycle Setting", *Journal of Political Economy*, 89, 1059-1085.

[18] Maddala, G. S. (1971): "The Use of Variance Components Models in Pooling Cross Section and Time Series Data", *Econometrica*, 39, 351-358.

[19] Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias", *Journal of Farm Economics*, 43, 44-56.

[20] Neyman, J. and E. L. Scott (1948): "Consistent Estimation from Partialy Consistent Observations", *Econometrica*, 16, 1-32.