1 Introduction

Comprehensive and good quality panel data over a long time period is rare, especially when the data is based on questionnaires. However, very often exists a series of cross section data. A often cited example is the household survey in Great Britain, where a long time no panels data existed but a large cross-section survey is carried out every year. Although this data is often regarded as inferior to true panel data, some classes of models can be consistently estimated using repeated cross section (RCS) data, by constructing so called pseudo panels.

Since it is impossible to track individual households over time, Deaton (1985) proposed to track cohorts through such data. A cohort is defined as a group with fixed membership, so that an individual is a member of exactly one cohort which is the same for all periods. Examples are age cohorts or cohorts based on sex or a combination of these variables. Successive surveys will generate successive random samples of individuals from these cohorts. Summary statistics from these time series can then be used to infer behavioral relationships for the cohort as if it were a panel. Panels constructed this way are called synthetic panels or pseudo panels.

Some disadvantages in the work with genuine panels do not arise in the work with pseudo panels. There is no attrition problem since new samples of individuals are drawn each year, and a further advantage is often the availability over a long time period.

2 The Errors in Variables Model

2.1 Estimation

Consider a linear model with individual fixed effects, such as

\[ y_{it} = \mathbf{x}_{it}'\beta + \theta_i + \epsilon_{it} \]  

(1)
The model follows the usual notation, \( \theta_i \) denote the individual effects, \( i \) and \( t \) indexes the individuals and time periods, respectively. The model is also sometimes referred to as Least Square Dummy Variables (LSDV) Model. If the individual effects \( \theta_i \) are uncorrelated with the explanatory variables, they can be estimated as random variables and for a consistent estimation of \( \beta \) be merged with the error-term \( \epsilon_{it} \). In this case the model can be estimated consistently by pooled OLS. If the individual effects \( \theta_i \) are correlated with the explanatory variables \( x_{it} \), this would lead to an inconsistent and biased estimator for \( \beta \).

When working with genuine panels this is solved by purging the individuals from their time averages, or equivalently by including dummy variables for each individual. The pseudo panel equivalence to the LSDV model in genuine panels is the following:

\[
\bar{y}_{ct} = \bar{x}_{ct}' \beta + \bar{\theta}_c + \bar{\epsilon}_{ct} \tag{2}
\]

\( c = 1, \ldots, C \) denote the cohorts and \( \bar{y}_{ct} \) and \( \bar{x}_{ct} \) are the cohort averages. However, the sample cohort means are error ridden estimates of the unobserved cohort population means and the unobserved true model is:

\[
y^*_{ct} = x^*_{ct}' \beta + \theta^*_c + \epsilon^*_{ct} \tag{3}
\]

Deaton (1985) assumes that the measurement errors have zero mean and are normally distributed

\[
y_{it} = y^*_{ct} + \zeta_{it} \\
x_{it} = x^*_{ct} + \eta_{it} \sim N \left( \begin{pmatrix} \zeta_{it} \\ \eta_{it} \end{pmatrix} \right), \begin{pmatrix} \sigma^2_{\zeta} & \sigma_{\zeta \eta} \\ \sigma_{\zeta \eta} & \Sigma_{\eta} \end{pmatrix} \tag{4}
\]

Then it follows for the measurement errors at cohort level that they are normally distributed with the true mean-values as mean, and variances which go to zero as the number of individuals in the cohort goes to infinity.

\[
\begin{pmatrix} \bar{y}_{ct} \\ \bar{x}_{ct} \end{pmatrix} \sim N \left( \begin{pmatrix} \bar{y}_{ct} \\ \bar{x}_{ct} \end{pmatrix}, \frac{1}{n_c} \begin{pmatrix} \sigma^2_{\zeta} & \sigma_{\zeta \eta} \\ \sigma_{\zeta \eta} & \Sigma_{\eta} \end{pmatrix} \right) \tag{5}
\]

Since the micro-data is used to construct the means, it can also be used to construct estimates of the variances and covariances of the sample means. Errors in variables estimators are used to derive consistent estimates of the population relationships. The variances and covariances of the measurement errors, \( \Sigma_{\eta} \) and \( \sigma^2_{\zeta} \) can be estimated from the micro-data. Once estimates for \( \Sigma_{\eta} \) and \( \sigma^2_{\zeta} \) are available, Deaton proposes the following error in variables estimator

\[
\hat{\beta} = (X'X - \hat{\Sigma}_{\eta})^{-1}(X'y - \hat{\sigma}_{\eta}) \tag{6}
\]

where the variance due to the measurement errors is subtracted from the moments or cross-product matrix. It is assumed assume that the time means have been already taken, so the estimator without being adjusted by the measurement error variance would be the standard FE-estimator.

The estimator \( \hat{\beta} \) is consistent if the number of observations \( TN \) tends to infinity.
3 A Generalisation of Deaton’s Approach

If the number of observations per cohort is large, it is tempting, and done in practice, to ignore the measurement error problem. Verbeek and Nijman (1992) analyse under which circumstances this can be a valid approach.

They show that the bias of the estimate will indeed be small if there is sufficient variation within the cohorts. Verbeek and Nijman (1992) assume that the regressor variables \( x_{it} \) are correlated with the cohort defining variable \( z_i \) in the following way:

\[
x_{it} = \mu_t + \gamma_t z_i + v_{it}
\]

(7)

The choice of larger cohorts will reduce the bias if either \( \mu_t \), \( \gamma_t \) or both vary with \( t \). The cohorts can be chosen smaller if this variation is large relative to the variance of \( v_{it} \).

Verbeek and Nijman (1993) also point out that there is a trade-off between the number of observations in a cohort and the number of cohorts in a panel. Since a decrease in the number of observations in a panel implies an increase in the variance of the FE-estimator, this results in a trade-off between bias and variance.

They show that consistency of the estimator proposed by Deaton requires that the number of time periods go to infinity. They introduce a new class of estimators with the standard FE-estimator and Deaton’s error in variables estimator as special cases. These new class of estimators are indexed by a parameter \( \alpha \in [0, 1] \) as follows:

\[
\hat{\beta}(\alpha) = (X'X - \alpha \Sigma)^{-1}(X'y - \alpha \sigma)
\]

(8)

where it is again assumed that the individuals or cohorts have been purged from their time mean in advance. The estimator proposed by Deaton is characterized by \( \alpha = 1 \). On the other hand the standard FE-estimator is characterized by \( \alpha = 0 \). The errors in variables estimator becomes consistent for finite \( T \) when \( \alpha \) is chosen to be \( \frac{T-1}{T} \). The trade-off between bias and variance implies that estimators where a smaller \( \alpha \) is chosen can perform better in terms of the mean squared error. Generally, the \( \alpha \) for the estimator which minimizes the mean squared error is smaller than \( \frac{T-1}{T} \).

4 An Instrumental Variables Interpretation

Moffitt (1993) proposes an alternative approach to estimation from repeated cross sections. He starts from the observation that grouping can be regarded as instrumental variables procedure. This is e.g. illustrated in Verbeek (1996).
First the individual effects $\theta_i$ are decomposed into a cohort effect $\theta_c$ and a individual deviation $v_i$:

$$\theta_i = \sum_{c=1}^{C} \theta_c^* d_{ci} + v_i$$  \hspace{1cm} (9)

where $d_{ci}$ are cohort dummies indicating that an individual $i$ is member of cohort $c$. We can then substitute (9) into (1) which yields

$$y_{it} = x_{it}\beta + \sum_{c=1}^{C} \theta_c d_{ci} + v_i + \epsilon_{it}$$  \hspace{1cm} (10)

It is not unlikely that the $x_{it}$ and $v_i$ are correlated, which would lead to inconsistent estimators using OLS. If instruments for $x_{it}$ can be found which are uncorrelated with $v_i$ and $\epsilon_{it}$, an instrumental variables estimator produces a consistent estimator for $\beta$ and $\theta_c$. Defining dummies $D_{st}$ for each time point, which are 1 if $s = t$ and otherwise 0, they can, interacted with the cohort dummies, be used as instruments for $x_{it}$. These dummies interacted with the cohort dummies can serve as linear predictor for $x_{it}$:

$$x_{it} = \sum_{c=1}^{C} \sum_{s=1}^{T} \gamma_{1,ct} d_{ci} D_{st} + \sum_{c=1}^{C} \gamma_{2,ct} d_{ci} + \omega_{it}$$  \hspace{1cm} (11)

Some restriction have to be made to avoid multicollinearity, e.g that the deviation from the cohort means sum up to 0. A linear predictor for $x_{it}$ is given by $\hat{x}_{it} = \bar{x}_{ct}$, which is the average value of $x$ in cohort $c$ at time $t$. The resulting IV estimator is given by

$$\hat{\beta}_{IV} = \left( \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)^2 \right)^{-1} \left( \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c) \right)$$  \hspace{1cm} (12)

which is just another way of writing the FE-estimator.

Moffitt (1993) presents a more general class of IV estimators. Instead of using cohort-dummies, time dummies and their interaction dummies, Moffitt’s (1993) approach allows a more parsimonious parametrization. He starts by the linear projection of $\theta_i$ on a set of time invariant variables $Z_i$, including not only cohort but e.g also residential location, years of schooling, . . . . The $x_{it}$ are projected on time invariant variables $z_i$ and time varying variables, denoted by $w_{it}$. The linear projection upon which the IV method is based can be written as:

$$x_{it} = \delta_1 w_{it} + \delta_2 z_i + u_{it}$$  \hspace{1cm} (13)

$$\theta_i = \gamma z_i$$  \hspace{1cm} (14)

The instrumental variables interpretation allows a whole new range of estimators. But as Verbeek (1996) remark, while the FE-estimator is consistent as the number of individuals per cohort tends to infinity, this is not guaranteed for estimators in the class suggested by Moffitt. They are consistent if the instruments for $x_{it}$ vary with $i$ conditional on the instruments for $\theta_i$. Having instruments for individual time series which do not exhibit individual variation might lead to inaccurate estimators.
References


