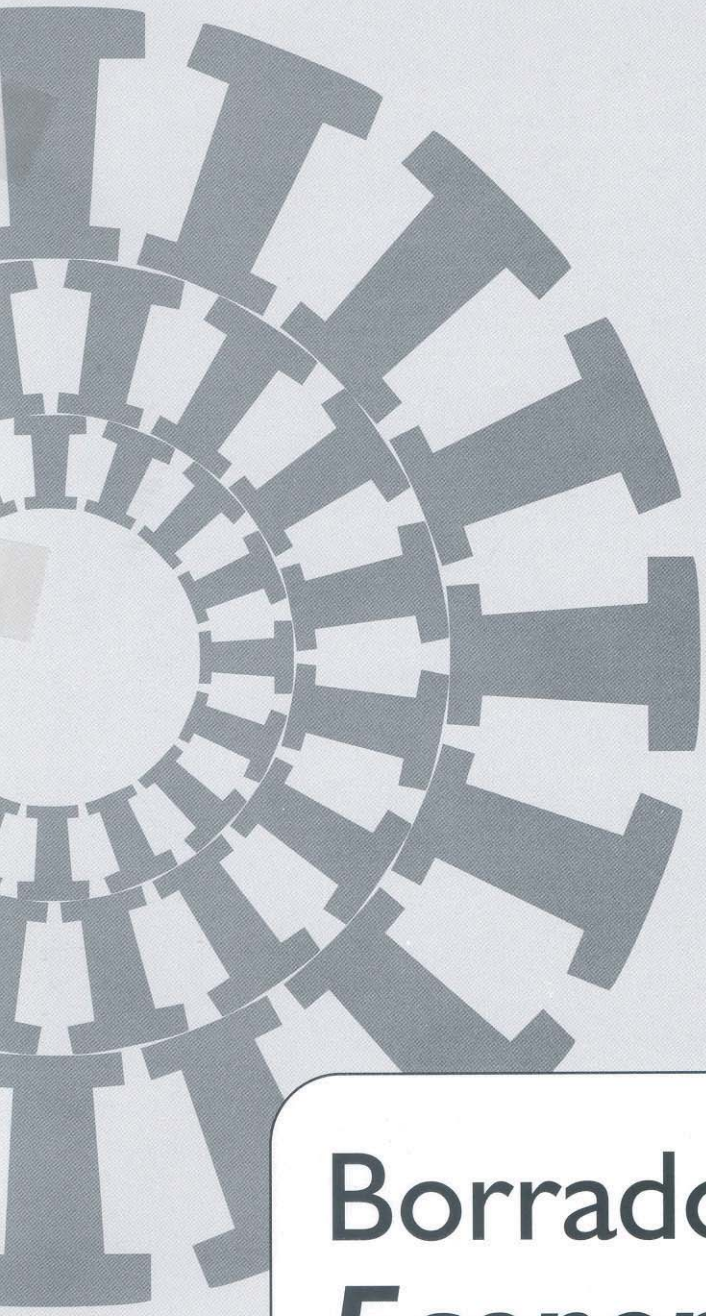


UNIVERSIDAD ICESI

Facultad de Ciencias Administrativas y Económicas

Borradores de *Economía y Finanzas*

Testing for sample selection bias in pseudo panels: Theory and Monte Carlo

Por:
Jhon James Mora
Juan Muro

No. 10, Marzo 2007

BORRADORES DE ECONOMÍA Y FINANZAS

Editor

Jhon James Mora

Jefe, Departamento de Economía

jjmora@icesi.edu.co

Asistente de edición

Stephanie Vergara Rojas

Gestión editorial

Departamento de Economía – Universidad ICESI

Contenido:

1. Introduction 3
2. Cohort data and selectivity bias in a cross-section model. 5
3. Pseudo panel data and selectivity bias 10
4. A sample selection bias correction term in Pseudo Panel Data. 15
5. A sample selection bias test in the IV form....17
6. Conclusions....22
7. References....24

24 páginas

ISSN 1900-1568
Primera edición, marzo de 2007

TESTING FOR SAMPLE SELECTION BIAS IN PSEUDO PANELS: THEORY AND MONTE CARLO

Jhon James Mora[1]

Juan Muro[2]

Abstract: Sample selection bias is commonly used in economic models based on micro data. Despite the continuous generalization of panel data surveys, most countries still collect microeconomic information on the behavior of economic agents by means of repeated independent and representative cross-sections. This paper discusses a simple testing procedure for sample selection bias in pseudo panels. In the context of conditional mean independence panel data models we describe a pseudo panel model in which under convenient expansion of the original specification with a selectivity bias correction term the method allows us to use a Wald test of H_0: $\rho=0$ as a test of the null hypothesis of absence of sample selection bias. We show that the proposed selection bias correction term is proportional to Inverse Mills ratio with an argument equal to the "normit" of a consistent estimation of the observed proportion of individuals in each cohort. This finding can be considered a cohort counterpart of Heckman's selectivity bias correction for the individual case and generalizes to some extent previous existing results in the empirical labour literature. Monte Carlo analysis shows the test does not reject the null for fixed T at a 5% significance level in finite samples and increases its power when utilizing cohort size corrections as suggested by Deaton (1985). As a "side effect" our method enables us to make a consistent estimation of the pseudo panel parameters under rejection of the null.

Key Words: Repeated Cross-section Models, Pseudo Panels, Selectivity Bias Testing, Discrete Analysis with Grouped Data, Monte Carlo Methods.

JEl Classifications: C23; C52

[1] Universidad Icesi y Universidad de Alcalá. e-mail: jjmora@icesi.edu.co

[2] Universidad de Alcalá. e-mail: juan.muro@uah.es

1. Introduction

Despite the continuous generalization of panel data surveys, most countries still collect microeconomic information on the behavior of economic agents by means of repeated independent and representative cross-sections. The current pseudo panel analysis starts with the seminal paper of Deaton (1985) who establishes that individual data can be replaced with cohort data with measurement error. Moffit (1991, 1993) introduces a consistent instrumental variable (IV) estimator for pseudo panel models using cohort dummies as instruments.

Sample selection bias is common in economic models based on micro data. Since Heckman (1979) selectivity bias treatment has been extended to panel data models by, among others, Wooldridge (1995), Kyriazidou (1998), Vella y Verbeek (1999), Rochina-Barrachina (1999) and Lee (2001) [see Jensen, Rosholm y Verter (2002) for a good survey of the literature]. Discussing sample selection bias in pseudo panels, however, is an unfinished task. Traditionally, empirical labour literature utilizes influential papers by Gronau (1974) and Lewis (1974) and eliminates selectivity bias by means of a correction term proportional to Mills inverse ratio with an argument equal to the inverse normal cumulative distribution function (normit) of the proportion of individuals observed in each cohort. Although selectivity analysis with grouped data is prior to Heckman´s contribution for the individual case, the connection between them remains unclear. The question has been formally presented in Moscarini and Vella (2002) from the perspective of an occupational mobility model with sample selection in which mobility and labour market participation equation errors are correlated. Some unanswered questions arise from Moscarini's and Vella's (2002) work. First of all, they do not discuss the presence of measurement errors in pseudo panel variables and consequently how to deal with in the line of Deaton (1985) or Moffitt (1991, 1993, 2007). Secondly, the selection variable in their model

is a pseudo panel variable as well and therefore with associated measurement error. Finally, given that in pseudo panel data we observe different individuals every period, we will obtain inconsistent estimators unless a set of assumptions on the selection process is established.

This paper presents a testing procedure for selectivity bias in pseudo panels. In the context of conditional mean independence panel data models we describe a pseudo panel model in which under convenient expansion of the original specification with a selection bias correction term the method allows us to use a Wald test of H_0: $\rho=0$ as a test of the null hypothesis of absence of sample selection bias. We show that the proposed selection bias correction term is proportional to Inverse Mills ratio of the normit of a consistent estimation of the observed proportion of individuals in each cohort. This finding can be considered a cohort counterpart of Heckman's selectivity bias correction term for the individual case and generalizes to some extent previous existing results in empirical labour literature. Monte Carlo analysis shows that the test does not reject the null for fixed T at a 5% significance level in finite samples and increases its power when utilizing cohort size corrections as suggested by Deaton (1985). As a "side effect" our method enables us to make a consistent estimation of the pseudo panel parameters under rejection of the null.

The paper is structured as follows: Section 2 provides a review of the consistent estimation of a cross-section grouped data model with selectivity bias. Section 3 discusses the consistency for pseudo panel IV estimators in presence of sample selection bias. In section 4 we introduce a selectivity bias correction term for pseudo panel models. In section 5 we propose a simple test for selectivity bias in pseudo panels and perform a Monte Carlo simulation to assess the power of the test. Finally, the conclusions are presented in section 6.

2. Cohort data and selectivity bias in a cross-section model.

In this section we review some results related with the consistent estimation of a cross-section model with grouped data and sample selection bias. The presentation ought to be formal, but for the moment it is limited to a mere collection of ideas linked by the principle of analogy.

Let us start with a cross-section model with individual data and sample selection bias. Let the model be:

$y_i^* = x'_i \beta + u_i$; $i = 1, \ldots, N$, (1)

$s^*_i = z'_i \gamma + v_i$; $s_i = 1[s^*_i > 0]$, (2)

Whereas, y_i^* denotes an interest variable in population model, i.e. over all observations. In (1) and (2) equation s_i is a selection process; u_i and v_i are usual errors.

As is well known, Heckman (1979), a consistent estimation of the equation of primary interest in (1) can be reached by ordinary least squares (OLS) by incorporating an additional selectivity bias correction term to (1). This term is:

$E(u_i | x_i, s^*_i > 0) \equiv E(u_i | x_i, s_i = 1) = E(u_i | x_i, z'_i \gamma + v_i > 0)$.

The final result under the assumption of joint normality of u_i and v_i with correlation ρ is that the selectivity correction term is proportional to Mills inverse ratio with argument $z'_i \gamma$, i.e.

$E(u_i | x_i, z'_i \gamma + v_i > 0) \propto \Phi(z'_i \gamma) / \Phi(z'_i \gamma)$,

where Φ (.) and Φ (.) are standard normal pdf and cumulative distribution functions, respectively. Note that in the individual case:

$$\text{Prob}\,(s^*_i > 0) \equiv \text{Prob}\,(s_i = 1) = \Phi\,(z'_i \gamma).$$

Then under normality assumption:

$$\Phi^{-1}[\Phi\,(z'_i \gamma)] = z'_i \gamma = \Phi^{-1}[\text{Prob}\,(s^*_i > 0)].$$

Consequently, in the individual context the argument of the Mills inverse ratio is the inverse standard normal cumulative distribution function with an argument equal to the probability associated with the observational rule ($s^*_i > 0$).

Let us continue with a cross-section model with grouped or cohort data and sample selection bias. Let the model be:

$$y_c^* = x'_c \beta + u_c \; ; \; c = 1, \ldots, C, \qquad (3)$$

$$s^*_c = z'_c \gamma + v_c; \qquad (4)$$

Where y_c^* denotes an interest variable in a cohort model. We observe that c runs for cohorts (C) and at this moment the model specification has a selection rule, but lacks an observational rule similar to the individual case. The model is the expression in cohorts of taking expectations in (1), (2) conditional to $i \in c$ (in random terms $g_i \in I_c$).

We would like to show that with cohort data the selectivity bias correction term is proportional to:

$$E(u_i | x_i, z'_i \gamma + v_i > 0, g_i \in I_c) \propto \phi(\Phi^{-1}(P_c)) / \Phi(\Phi^{-1}(P_c)), \quad (5)$$

Mills inverse ratio, the argument of which $\Phi^{-1}(P_c)$ is the inverse standard normal cumulative distribution function or normit of the observed proportion of individuals in cohort c (size of c, N_c, tends to infinity).

In the cohort case:

$$P_c = \text{Prob}(s^*_i > 0, g_i \in I_c),$$

the observed proportion is the probability of the observational rule. Via analogy it seems clear that in the cohort case the argument of Mills inverse ratio will be as above the inverse standard normal cumulative distribution function with an argument equal to the probability of $s^*_i > 0, g_i \in I_c$, i.e. $\Phi^{-1}(P_c)$

This selectivity bias correction term was first introduced by Gronau (1974) and Lewis (1974) in an analysis of wage comparisons with grouped or cohort data. Translating Gronau-Lewis proposal into our own econometric language we can say that for them the observed participation rate, the result of the observational rule at cohort level, is only determined by cohort. Formally,

$$\Phi^{-1}(P_c) = \Sigma \alpha_c D_c + v_c, \; c=1, 2, \ldots\ldots C, \quad (6)$$

where D_c are cohort dummies. In other words, in their analysis the main and only source of variation for the observed participation rate was cohort. So in order to avoid selectivity biases they recommended adding a new variable to the equation of primary interest whose data were Mills inverse ratio with an argument equal to the inverse standard normal cumulative distribution function

of the observed participation rate. The Gronau (1974) and Lewis (1974) suggestion has often been used in the empirical labour literature, see for example Blundell et al. (1998).

Generalizing this finding we know that the cohort counterpart to the individual selection equation is (4). When the observational rule applies we obtain the following

$$s^*_c \equiv \Phi^{-1}(P_c) = z'_c \gamma + v_c. \qquad (7)$$

The above expression emphasizes that the selection rule does not eliminate in general any cohort and what is actually observed is the proportion of individuals that are in deed observed in each cohort. As can be seen, (7) generalizes (6) and allows the inclusion of a set of determinants of the participation rate different from a group of cohort dummies.

Many arguments can be given to support the idea that improving the specification of equation (6) will lead to better estimates of the equation of interest. To say the least in the empirical labour literature is usual to assume that variables such as age, education, and household characteristics, among others, play an important role among the determinants of the participation rate and therefore must be included in the specification of the selection equation. Needless to say that the equivalence between (6) and (7) can be achieved through a thorough definition of cohorts so that each cohort only contains homogeneous individuals in terms of the complete set of determinants of the participation rate. This argument is theoretically unbeatable, but empirically weak because cohorts are usually defined in terms of a small set of variables just to preserve the desired size.

Finally, in the line of Gronau (1974) and Lewis (1974) to obtain a consistent estimation of the equation of interest in (3) we propose a two step method. In the first step we consistently estimate (7) (by means for instance of OLS with heteroskedasticity correction or maximum likelihood (ML)) and in a second step we carry out an OLS estimation of the equation of interest (3) augmented with an additional selectivity bias correction term of the form:

$$\Phi (z'_c \gamma)/ \Phi (z'_c \gamma) \equiv \Phi (\Phi^{-1} (P_c))/ \Phi (\Phi^{-1} (P_c)) \quad (7a)$$

evaluated at consistent estimates obtained in the first step. A test of the presence of selectivity bias in (3) is a test of significance of the parameter of the selectivity bias correction variable in the augmented regression that can be performed in whatever usual ways.

3. Pseudo panel data and selectivity bias

Let us consider the pseudo panel data model with sample selection:

$$y_{i(t),t} = x'_{i(t),t}\beta + \alpha_{i(t)} + u_{i(t),t} \quad ; \quad t = 1,\ldots,T \quad ; i_{(t)} = 1, \ldots, N_t \,. \tag{8}$$

$$s^*_{i(t),t} = z'_{i(t),t}\gamma + \mu_{i(t)} + v_{i(t),t}\,; \qquad s_{i(t),t} = 1[s^*_{i(t),t}>0]. \tag{9}$$

Where (8) is the equation of primary interest and (9) contains both a selection equation and an observational rule. We denote $y_{i(t),t}$ as an interest variable in repeated cross section model with measurement error.[3] On other hand, $y_{i(t),t}$ is only observed when $s_{i(t),t}$ =1 and β, γ are parameter vectors; $x_{i(t),t}$, $z_{i(t),t}$ are covariates; $\alpha_{i(t)}$,$\mu_{i(t)}$ are individual effects in t; $u_{i(t),t}$, $v_{i(t),t}$ are idiosyncratic errors; i run for individuals. Our data consist of a time series of independent cross-sections so we can only observe the same individual in one period of time.

When individual effects, $\alpha_{i(t)}$, are uncorrelated with explanatory variables, $x_{i(t),t,}$ equation in (8) can be estimated by pooling ordinary least squares (OLS) considering $\alpha_{i(t)} + u_{i(t),t}$ as a compound error even though the variance of $\alpha_{i(t)}$ is not identified. However, in most situations individual effects are correlated with explanatory variables. So considering $\alpha_{i(t)}$ as a random component following a specific probability distribution leads to inconsistent estimation of the parameters in (8). This inconsistency can be solved regarding $\alpha_{i(t)}$ as an unknown parameter.

Deaton (1985) suggests using cohorts to obtain consistent estimations of β in (8) when we have repeated and independent cross-sections data even in the case of correlation between individual

[3] That is, over all individuals in a specific cohort.

effects and explanatory variables. Moffitt (1993) recommends using IV and decomposes the individual effect $\alpha_{i(t)}$ in a cohort effect α^*_c plus an individual deviation $\tau_{i(t)}$. Thus:

$$\alpha_{i(t)} = \sum_{c=1}^{C} d'_c \alpha_c + \tau_{i(t)} \ , \qquad (10)$$

Where d_c is equal to 1 if individual i belongs to cohort c and 0 otherwise. Substituting (10) in (8) we obtain

$$y_{i(t),t} = x'_{i(t),t}\beta + \sum_{c=1}^{C} d'_c \alpha_c + \tau_{i(t)} + \mu_{i(t),t} \ ; \ \ t = 1,....,T. \qquad (11)$$

In equation (11) provided we have a set of instruments for $x_{i(t),t}$ uncorrelated with $\nu_{i(t),t}$ y $\mu_{i(t),t}$, the IV estimator is a consistent estimator for β y α^*_c. A set of temporary dummies, $D_{s,t} = 1$ if s = t and 0 otherwise, and interactions with cohort dummies can be used as instruments for $x_{i(t),t}$,. Thus, the reduced form linear predictor will be

$$x_{i(t),t} = \sum_{c=1}^{C} \sum_{t=1}^{T} D'_{s,t}\varsigma_1 + \sum_{c=1}^{C} d'_{c,t}\varsigma_2 + \omega_{i(t),t} \ , \qquad (12)$$

Where $\omega_{i(t),t}$ is an error vector. The lineal predictor for $x_{i(t),t}$ is $\hat{x}_{i(t),t} = \bar{x}_{ct}$ the average of $x_{i(t),t}$ in cohort c and period t. Then the IV estimator of β is:

$$\hat{\beta}_{IV} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c) \right) . \qquad (13)$$

Consistency conditions for the estimator in (13) imply that instruments for $x_{i(t),t}$ must vary with t and are asymptotically uncorrelated with $\nu_{i(t)}$ y $\mu_{i(t),t}$, Verbeek (1996).

When our sample comes from incidental truncation estimators the pseudo panels are in general inconsistent due to the presence of sample selection bias, Heckman (1979). Note that in the case of identical sample selection processes for all individuals across periods, the fixed effect estimator for the pseudo panel would also eliminate selectivity bias. However, this assumption is very difficult to maintain. Additionally, the presence of unobserved individual heterogeneity in the selection process would lead to inconsistencies unless this heterogeneity is dealt with in an appropriate way. In particular, unobservable effects and selectivity bias could be removed through differencing, but this method is unfeasible in pseudo panels.

In presence of sample selection bias let $s^*_{i(t),t}$ be a selection index so that $\{y_{i(t),t}, x_{i(t),t}\}$ are only observed when $s_{i(t),t}$ equals 1. Then the IV estimator in (13) becomes

$$\hat{\beta}_{IV} = \left[\left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_{c}) \right) \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_{c})' \right) \right]^{-1} \times \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_{c}) \right) \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{y}_{ct} - \bar{y}_{c}) \right) \quad (14)$$

Let us introduce a set of assumptions (AS1-AS2).

AS1: $$p \lim_{N_c \to \infty} \frac{1}{NT} \sum_{i(t)=1}^{N} \sum_{t=1}^{T} s_{i(t),t} \mu_{i(t),t} = 0 \quad (15)$$

where N_c is the number of individuals in each cohort. So:

$$\lim_{N_c\to\infty}\frac{1}{NT}\left[E\left(\sum_{i(t)=1}^{N}\sum_{t=1}^{T}s_{i(t),t}\mu_{i(t),t}\right)\right]=\lim_{N_c\to\infty}\frac{1}{NT}\left[\sum_{i(t)=1}^{N}\sum_{t=1}^{T}E\left(s_{i(t),t}\mu_{i(t),t}\right)\right]=0 \tag{16}$$

AS2: $$p\lim_{N_c\to\infty}\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T}s_{i(t),t}\tau_{i(t),t}=0 \tag{17}$$

Consequently,

$$\lim_{N_c\to\infty}\frac{1}{NT}\left[E\left(\sum_{i(t)=1}^{N}\sum_{t=1}^{T}s_{i(t),t}\tau_{i(t),t}\right)\right]=\lim_{N_c\to\infty}\frac{1}{NT}\left[\sum_{i(t)=1}^{N}\sum_{t=1}^{T}E\left(s_{i(t),t}\tau_{i(t),t}\right)\right]=0 \tag{18}$$

It is worth noting that AS2 holds true because of the fact that the deviation of heterogeneity with respect to the cohort is independent from the selection process itself. However, hold AS1 is more disputable if the individuals are not selected at random.

Proposition 3.1

Under AS1 and AS2 the estimator $\hat{\beta}_{IV}$ is consistent for fixed T and $N_c\to\infty$.

Proof:

$$\begin{aligned}\hat{\beta}_{IV} &= \left[\left(\sum_{c=1}^{C}\sum_{t=1}^{T}s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})\right)\left(\sum_{c=1}^{C}\sum_{t=1}^{T}s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})'\right)\right]^{-1}\\ &\times\left(\sum_{c=1}^{C}\sum_{t=1}^{T}s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})\right)\left(\sum_{c=1}^{C}\sum_{t=1}^{T}s_{i(t),t}(\bar{y}_{ct}-\bar{y}_{c})\right)\end{aligned} \tag{19}$$

$$
\begin{aligned}
p\lim \hat{\beta}_{IV} = \beta + p\lim\left[\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})\right)\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})'\right)\right]^{-1} \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})'\right) \\
\times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}\tau_{i(t),t}\right) + p\lim\left[\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})\right)\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})'\right)\right]^{-1} \\
\times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_{c})'\right) \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}\mu_{i(t),t}\right)
\end{aligned}
\tag{20}
$$

So:

$$
p\lim \hat{\beta}_{IV} = \beta \tag{21}
$$

as we want to show.

4. A sample selection bias correction term in Pseudo Panel Data.

In the pseudo panel case with selectivity bias the cohort expression for the equation (8) will be as follows:

$$E(y_{i(t),t} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c)=$$

$$E(x'_{i(t),t}\,\beta + \tau_{i(t)} + \mu_{i(t),t} \mid\ x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c)=$$

$$E(x'_{i(t),t}\,\beta \mid x_{i(t),t}, s_{i(t),t}=1, g_{i(t)} \in I_c)+E(\tau_{i(t)} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c)+E(\mu_{i(t),t} \mid x_{i(t),t}, s_{i(t),t}=1, g_{i(t)} \in I_c) \quad (22)$$

In equation (22) $g_{i(t)} \in I_c$ shows that observation i(t) in the appropriate cross section belongs to a specific cohort. The solutions for pseudo panel data show that the direct procedure for the first term in equation (22) implies the use of the sample mean of the variables in the respective cohorts. By AS2 the second term becomes zero while the deviation of the cohort is independent from the selection process. There is, however, no guarantee that the last term equals zero, which shows that the estimator is inconsistent when there is a potential selection bias.

Because the selection process does not affect the presence or absence of a cohort in a specific cross section, cohorts will comprise a set of different individuals in each repeated cross section, and the presence of different individuals in each cross-section is independent from the incidental truncation process. Therefore, a random selection of representative samples of each sub-population of cohorts will contain different individuals in each cross section. This makes it necessary to find an expression that allows inferring the behavior of a cohort based on the behavior of different individuals in the cohort. Thus, the last expression in equation (22) is

$$E(\mu_{i(t)} | x_{i(t),t}, s_{i(t),t}=1, g_{i(t)} \epsilon I_c) = E(R_{i(t)} | g_{i(t)} \epsilon I_c) \quad (23)$$

In equation (23) above, $R_{i(t)}$ is Mills inverse ratio which shows the transformation of individual results into cohort results. It is worth noting that if the nature of the selection process is known, then it is possible to use individual parameters (estimated for the selection process) and apply them to the means of the cohort to obtain a "selection indicator" for each cohort.

To evaluate expression in (23) we follow the procedure reviewed in section 1 for the cross-section model. Instead of integrating out the individual Mills inverse ratio for all the observed individuals in each cohort we calculate Mills inverse ratio for the normit of a consistent estimation of the observed proportion of individuals in each cohort. An additional difference in the pseudo panel case is that we have to condition on all the results of the selection process and so the consistent estimation of the proportion must be obtained from a consistent pseudo panel estimation of the selection equation.

5. A sample selection bias test in the IV form

One of the ways to identify the existence of selection biases consists of modeling (23), following the work of Heckman (1979), and contrasting the hypothesis that the expectation is equal to zero. [4] Let us assume the following selection process with instrumental variables following Moffit's (1993) work,

$$s_{i(t),t} = 1\,[s_{i(t),t}{}^{*} > 0] = 1[r'_{i(t),t}\,\beta + \eta_{i(t)} + e_{i(t),t} > 0] \qquad (24)$$

In equation (24) $r_{i(t),t}$ is always observed unlike $\{y_{i(t),t}\,,\, x_{i(t),t}\}$, which are observed only when $s_{i(t),t}$ equals 1. On the other hand, $\eta_{i(t)}$ represents non-observable individual heterogeneity and $e_{i(t),t}$ is the error. And $1[\bullet]$ is the indicator function. If we break down individual heterogeneity into a cohort effect and a deviation, the following result is achieved:

$$\eta_{i(t)} = m'_{i(t),t}\,\zeta_1 + \varphi_{i(t)} \qquad (25)$$

If we substitute (25) in (24), we will obtain:

$$s_{i(t),t} = 1[r'_{i(t),t}\,\beta + m'_{i(t),t}\,\zeta_1 + \varphi_{i(t)} + e_{i(t),t} > 0] \qquad (26)$$

$$r_{i(t),t} = m'_{i(t),t}\,\zeta_1 + z'_{i(t),t}\,\zeta_2 + \omega_{i(t),t} \qquad (27)$$

In equation (27) a linear projection of $r_{i(t),t}$ is performed on time-invariant variables such as cohorts and a set of $z_{i(t),t}$ additional variables and $\xi_{i(t),t} = \omega_{i(t),t} + e_{i(t),t}$. Following the work of Heckman

[4] A contrast following Nijman-Verbeek (1992) or Woldridge (2005), in which lagging or leading values of $s_{i(t),t}$ are added to the main equation, will not work as long as $s_{i(t),t-1} \neq s_{i(t),t} \neq s_{i(t),t+1}$ because the individual is observed only once.

(1979) and Wooldridge (1995), let us assume that { $\mu_{i(t),t}$, $\xi_{i(t),t}$ } is independent from {$\alpha_{i(t),t}$, $m_{i(t),t}$ }. Thus, if $E(\mu_{i(t),t} | \xi_{i(t),t})$ is linear, then

$$E(\mu_{i(t),t} | \tau_{i(t),t} , s_{i(t),t}) = \rho E(\xi_{i(t),t} | \tau_{i(t),t} , s_{i(t),t}) = \rho E(\xi_{i(t),t} | s_{i(t),t}) \quad (28)$$

And the main equation is thus rewritten accordingly

$$y_{i(t),t} = x'_{i(t),t}\beta + \tau_{i(t)} + E(\xi_{i(t),t} | s_{i(t),t})\rho + \psi_{i(t),t} \quad ; \; E(\psi_{i(t),t} | \tau_{i(t)}, x_{i(t),t}, s_{i(t),t}) = 0 \quad (29)$$

Consequently, if $E(\xi_{i(t),t} | s_{i(t),t})$ is known, then a contrast about the existence of selection biases will involve contrasting the hypothesis of a lack of significance of ρ in (29). It must be noted that, because of the existence of non-observable individual heterogeneity in the selection equation, if this is not properly addressed, one could conclude that the existence of a selection bias may be due to the existence of some correlation between non-observed individual heterogeneity and some explanatory variable.

5.1 A Simple selectivity bias testing procedure.

The procedure described below, which contrasts the existence of selection biases, is valid under the null hypothesis of a lack of selection biases. We need the following assumption,

AS3. $E(\xi_{c,t} | s_{c,t}) = E(\xi_{i(t),t} | s_{i(t),t})$ when $s_{i(t),t} = r'_{c,t}\beta + m'_{c,t}\zeta + \varphi_{i(t)} + \xi_{i(t),t}$ (30)

This assumption implies when we instrument (24) with cohort variables then the expected probability over individuals is an expected probability over cohort. Then if AS3 is valid, the methodology for the selectivity bias could be,

1) Using an iv-probit with cohorts as instruments to estimates $E(\xi_{i(t),t} \mid s_{i(t),t})$.
2) Determining Mills inverse ratio, $\hat{\lambda}_{i(t),t}$, using the previous equation.
3) For the sample in which $s_{i(t),t} = 1$, estimating (29) by instrumental variables, by replacing $E(\xi_{i(t),t} \mid s_{i(t),t})$ with $\hat{\lambda}_{i(t),t}$.
4) Hypothesis Ho: $\rho = 0$ may then be compared against the value of t or the p-value may be used with a certain level of significance.

5.2 Power of the IV-test

The following is a description of the Monte Carlo experiment, which was conducted to investigate the power of the contrast proposed in section 5.1 above. First, a set of individual series was generated in each period, including cohort dummies used to keep track of individuals over time [see Vella and Verbeek (2005), Girma (2001), and Verbeek and Nijman (1993)]. Thus, the selection equation was generated as shown below:

$$r_{i(t),t} = f_{i(t),t} + c_{i(t),t} + \omega_{i(t),t} \quad (31)$$

$$s_{i(t),t} = 1[\, r_{i(t),t} + c_{i(t),t} + \eta_{i(t),t} > 0] \quad (32)$$

In equation (32) $f_{i(t),t}$ was generated at random from a normal distribution; $c_{i(t),t}$ consists of 10 dummies of cohorts with identical probability; $\omega_{i(t),t}$ was generated at random from a normal distribution; and $\eta_{i(t),t}$ was generated at random from a uniform normal distribution N [0,1]. The individuals were selected from a percentile-based distribution of (32). For example, when 50% of the individuals are selected, (32) was divided with the same mass of probability (32). The main equation was generated as follows:

$$X_{i(t),t} = d_{i(t),t} + c_{i(t),t} + \xi_{i(t),t} \quad (33)$$

$$Y_{i(t),t} = X_{i(t),t} + \varphi_{i(t),t} \quad (34)$$

The observations about (33) and (34) were made in the same way as (31) and (32). Then, the linear projection of (33) on the cohorts was performed, and Inverse Mills ratio was calculated. The latter was then incorporated into the main regression, equation (29). Therefore, this provides an analysis of the power of contrast proposed based on the null hypothesis that ρ equals zero. The corresponding results are listed below:

Table 1. Monte Carlo simulations $Y_{i(t),t} = \beta' X_{i(t),t} + C_{i(t)} + \rho' \lambda_{i(t),t} + \psi_{i(t),t}$.

$S_{i(t)}$ / T	5			7			10		
	β	Sd_β	Power	β	Sd_β	Power	β	Sd_β	Power
10%	0.9999718	0.0105441	0.05	0.9998952	0.0089107	0.048	0.9998906	0.007455	0.04
30%	0.9999718	0.0105441	0.05	1.000186	0.0100972	0.049	1.000159	0.008456	0.048
50%	1.000047	0.0141582	0.053	0.9996328	0.0119637	0.051	1.000404	0.0100063	0.05

Note: Average values for β, Sd_β and ρ with 1,000 iterations

Table 1 above shows the results of the Monte Carlo simulation using 1,000 iterations, 10 cohorts, and 2,000 individuals. These results show with a significance level of 5% that the power of contrast for a fixed T increases as the number of individuals increases. It must also be noted that the power of contrast drops down to a percentage that is not greater than 5.3% when T=5. Therefore, the results show that, for a fixed selection size, as the period of time increases, the power of the contrast comes close to a significance level of 5%.

This is followed by weighting the values using the square root of the size of the cohort in order to determine the effect of the cohort on the power of the contrast:

Table 2. Monte Carlo simulations $Y_{i(t),t} = \beta' X_{i(t),t} + C_{i(t)} + \rho' \lambda_{i(t),t} + \psi_{i(t),t}$.

$S_{i(t)}$ / T	5				7				10			
	β	Sd_β	ρ	Power	β	Sd_β	ρ	Power	β	Sd_β	ρ	Power
10%	1.00	0.0291	0.0039	0.0	1.0	0.0245	-0.0079	0.0	0.99	0.0204	-0.0037	0.0
30%	1.00	0.0455	-0.002	0.0	1.0	0.0382	-0.0034	0.0	1.00	0.0319	-0.0018	0.0
50%	0.99	0.0499	0.0008	0.0	0.99	0.0421	-0.0030	0.0	1.00	0.035	0.00111	0.0

Note: Averages for β, $Sd_{\beta\text{ , and }}$ ρ using 1,000 iterations

The results listed in Table 2 above show that the power of the contrast increases when the values are weighted using the size of the cohort. They also show that the null hypothesis of a lack of selection bias is not ruled out on any significance level.

6. Conclusions

Moffitt's estimator (1991, 1993) of pseudo panel data is consistent when there are no selection biases. Since there is no apparent reason to believe that the selection process is time-invariant, then the presence of a selection bias leads to inconsistent estimators.

This paper discusses a simple testing procedure for sample selection bias in pseudo panels. In the context of conditional mean independence panel data models we describe a pseudo panel model in which under convenient expansion of the original specification with a selectivity bias correction term the method allows us to use a Wald test of H_0: $\rho=0$ as a test of the null hypothesis of absence of sample selection bias. We show that the proposed selection bias correction term is proportional to Inverse Mills ratio with an argument equal to the “normit” of a consistent estimation of the observed proportion of individuals in each cohort. This finding can be considered a cohort counterpart of Heckman’s selectivity bias correction for the individual case and generalizes to some extent previous existing results in the empirical labour literature.

On the other hand, this paper discusses the consistency issues of the pseudo panel estimator when instrumental variables are used in presence of selection biases. Following the above results our selection bias contrast implies the use of instrumental variables in the selection equation of the pseudo panel. The characteristics of this contrast are analyzed based on Monte Carlo simulations, using 1,000 iterations, 10 cohorts, and 2,000 individuals at three different time points, i.e. t=5, t=7, and t=10, and a selection bias of 90%, 70%, and 50%, respectively. The results show with a significance level of 5% that the power of contrast for a fixed T increases as the number of individuals increases. The results also show that, for a fixed selection size, as the period of time increases, the

power of the contrast comes close to a significance level of 5%. When the values are weighted using the square root of the size of the cohort, the power of the contrast will increase significantly.

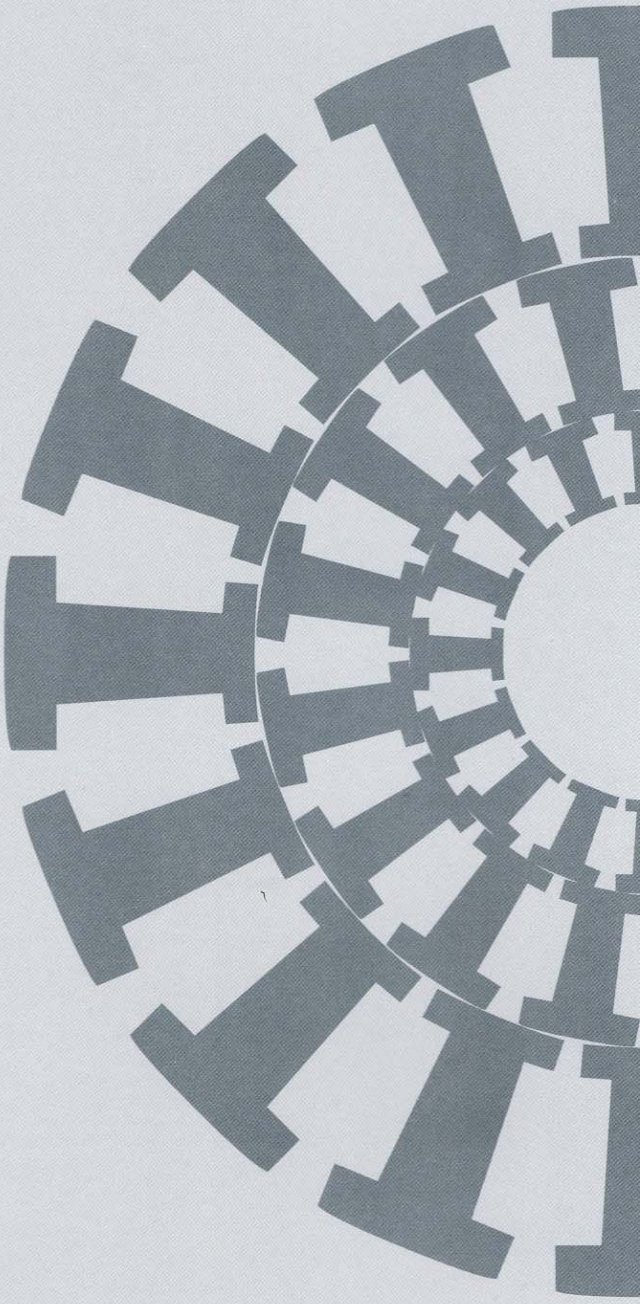
7. References

- Blundell, R., A. Duncan, C. Meghir (1998), “Estimating Labor Supply Responses Using Tax Reforms”, Econometrica, 66: 827-861.
- Deaton, A. (1985), “Panel data from time series of cross-sections”, *Journal of Econometrics*, 30: 109-126.
- Dustman, C., Rochina-Barrachina, M. (2000), “Selection Correction in Panel Data Models: An Application to Labour Supply and Wages,” Discussion Paper No. 162, IZA
- Gronau, R. (1974), “Wage Comparisons, A Selectivity Bias”, *Journal of Political Economy*, 82: 1119-1144.
- Heckman, J. (1979), “Sample selection bias as a Specification Error”, *Econometrica*, 47: 153-161.
- Jensen, P., Rosholm, Verner M. (2002), “A Comparison of Different Estimators for Panel Data Sample Selection Models”, University of AARHUS, W.P. No. 2002-1.
- Kyriazidou, E. (1998), “Estimation of a Panel Data Sample Selection model”, *Econometrica*, 65: 1335-1364.
- Lee, M.J. (2001), First-Difference Estimator for Panel Censored-Selection Models, *Economics Letters* 70: 43-49.
- Lewis, H.G. (1974), “Comments on Selectivity Biases in Wage Comparisons”, *Journal of Political Economy*, 82: 1145-1155.
- Moffitt, R. (1991), “Identification and estimation of Dynamic Models with a Time Series of Repeated Cross-Sections”, Brown University, Providence RI, mimeo.

- ----------(1993), “Identification and estimation of Dynamic Models with a Time Series of Repeated Cross-Sections”, *Journal of Econometrics* 59: 99-123.
- Moscarini, G., F. Vella. (2002), “Aggregate Worker Reallocation and Occupational Mobility in the U.S.:1971-2000”, IFS Working Papers, W02/18.
- Rochina-Barrachina, M.E. (1999), “A New Estimator for Panel Data Sample Selection Models”, *Annales d´Ėconomie et de Statistique*, 55/56:153-181.
- Ridder, G and R. Moffitt (2007), “The Econometrics of Data Combination” in Handbook of Econometrics, Vol 6 , Elsevier, Forthcoming
- Verbeek, M (1996), “Pseudo Panel Data”, in: L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data: Handbook of Theory and Applications*, Second Revised Edition, Kluwer Academic Publishers, Dordrecht, pp. 280-292.
- Vella, F., Verbeek, M (1999), “Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias”, *Journal of Econometrics* 90: 239-263
- ----------(2005), **“**Estimating Dynamic Models from Repeated Cross-Sections”. J*ournal of Econometrics*, 127(1): 83-102.
- Wooldridge, J.W. (1995), ”Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions”, *Journal of Econometrics*, 68:115-132.
- --------(2002), *Econometric Analysis of Cross Section and Panel Data*. The MIT press.

RESUMEN "BORRADORES DE ECONOMÍA Y FINANZAS"

1	Jhon J. Mora	El efecto de las características socio-económicas sobre la consistencia en la toma de decisiones: Un análisis experimental.	May-01
2	Julio C. Alonso	¿Crecer para exportar o exportar para crecer? El caso del Valle del Cauca.	Mar-05
3	Jhon J. Mora	La relación entre las herencias, regalos o loterías y la probabilidad de participar en el mercado laboral: EL caso de España, 1994-2000.	Jun-05
4	Julián Benavides	Concentración de la propiedad y desempeño contable: El caso latinoamericano.	Sep-05
5	Luís Berggrun	Price transmission dynamics between ADRD and their underlying foreign security: The case of Banco de Colombia S.A.- BANCOLOMBIA	Dic-05
6	Julio C. Alonso y Vanesa Montoya	Integración espacial del mercado de la papa en el Valle del Cauca: Dos aproximaciones diferentes, una misma conclusión	Mar-06
7	Jhon J. Mora	Datos de Panel en Probit Dinámicos	Jun-06
8	Julio C. Alonso yMauricio Arcos	Valor en Riesgo: evaluación del desempeño de diferentes metodologías para 7 países latinoamericanos	Ago-06
9	Mauricio Arcos y Julian Benavides	Efecto del ciclo de efectivo sobre la rentabilidad de las firmas colombianas	Dec-06
10	Jhon J. Mora y Juan Muro	Testing for sample selection bias in pseudo panels: Theory and Monte Carlo	Mar-07



UNIVERSIDAD
ICESI

Calle 18 No. 122 - 135 - Cali - Colombia
Tel. 555 2334 Ext. 419 - Fax 555 2345
http://www.icesi.edu.co/~econego/depto/

ISSN 1900-1568