Consistent Estimation in Pseudo Panels in the Presence of Selection Bias

Jhon James Mora Rodríguez
Universidad Icesi, Colombia
Juan Muro
Universidad de Alcalá, Spain

Abstract*:* Sample selection bias is common in economic models based on micro data. In the presence of selection bias the traditional estimators for pseudo panel data models are inconsistent. This paper discusses a method to achieve consistency in static pseudo panels in the presence of selection bias and a simple testing procedure for sample selection bias. We describe a pseudo panel model in which under convenient enlargement of the original specification with a selectivity bias correction term our procedure allows to test for sample selection bias. We show that in the line of Gronau (1974) and Lewis (1974) the proposed selection bias correction term is proportional to the inverse Mills ratio with 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 labor 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" we use the enlarged pseudo panel to provide a GMM consistent estimation of the pseudo panel parameters under rejection of the null.

JEL C23 C52
Keywords: Repeated Cross-section Models, Selectivity Bias Testing, Returns to Education.

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 (RCS). The current pseudo panel analysis starts with the seminal article 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 (1976, 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 labor literature utilizes influential articles by Gronau (1974) and Lewis (1974), hereafter G-L, and eliminates selectivity bias by means of a correction term proportional to the inverse Mills 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.

This article presents a testing procedure for selectivity bias in pseudo panels. 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 the absence of sample selection bias. We show that the proposed selection bias correction term is proportional to the 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 article 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. Selectivity bias in a cross-section model and in a repeated cross-section (RCS) model.

In this section we review some results related with the consistent estimation of a cross-section model with individual data and sample selection bias and in turn we analyze the repeated cross-section (RCS) model in the presence of selection bias.

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

$$y_i^* = x'_i \beta + u_i \; ; \; i = 1, \ldots, N, \tag{1}$$

$$s_i^* = z'_i \gamma + v_i; \qquad s_i = 1[s_i^* > 0], \tag{2}$$

$$y_i = y_i^* \text{ when } s_i = 1; \; y_i \text{ unobserved otherwise} \tag{3}$$

Where y_i^* is the variable of interest, s^*_i the selection, z explain s_i, and u_i, v_i are usual errors. Usual exclusion restrictions hold. As is well known, Heckman (1979), a consistent estimation of the equation of primary interest in (1) can be obtained by ordinary least squares (OLS) by adding a selectivity bias correction term in (1). This term is

$$E(u_i | x_i, s_i^* > 0) \equiv E(u_i | x_i, s_i = 1) = E(u_i | x_i, z'_i \gamma + v_i > 0). \tag{4}$$

The final result under the assumption of joint normality of u_i and v_i with correlation ρ (or a less restrictive assumption as $E(u_i | v_i) = \rho v_i$) is that the selectivity correction term is proportional to the inverse Mills ratio (IMR) with argument $z'_i \gamma$, i.e.

$$E(u_i | x_i, z'_i \gamma + v_i > 0) \propto \Phi(z'_i \gamma) / \Phi(z'_i \gamma), \tag{5}$$

Where $\Phi(.)$ and $\Phi(.)$ 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

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

And (5) can be rewritten as

$$E(u_i| x_i, s_i^* > 0) \equiv E(u_i| x_i, s_i =1) \propto \Phi(\Phi^{-1}[Prob(s_i =1)])/ \Phi(\Phi^{-1}[Prob(s_i =1)]). \quad (6)$$

Hence with individual data the argument of the IMR is the inverse standard normal cumulative distribution function or normit function of the probability associated with the observational rule ($s_i^* > 0$). This is a standard result of the statistical literature.

Let us now continue with a repeated cross-section (RCS) model with sample selection bias. The sample model for individual i and time t is

$$y_{i(t),t} = x'_{i(t),t}\beta + u_{i(t),t}; \quad i = 1, \ldots, N_t; t = 1, 2.....T; \; y_{i(t),t} \text{ is only observed when } s_{i(t),t} =1, \quad (7)$$

where subscript (t) means different individuals are observed in each time period t. To simplify notation we will drop subscript (t) hereafter. As we observe different individuals in a RCS model we use cohort dummies as matching instruments. Taking expectations in (1) we get the cohort population model

$$E(y_{it}^*|X_{it}, g_i \in I_c) = E(x'_{it}|X_{it}, g_i \in I_c)\beta + E(u_{it}|X_{it}, g_i \in I_c); i= 1,.,N_t; t= 1,.,T; \; c=1, 2....C, \quad (8)$$

where $g_i \in I_c$ denotes that individual i belongs to a specific cohort c. The cohort regression in the absence of selection bias (8) can be used as an errors-in-variable estimating equation taking sample cohort-means as population cohort-means subject to errors, Deaton (1985). In the presence of selection bias however the relevant equation is (7) and taking expectations

$$E(y_{it}|Z_{it}, s_{it}=1 |g_i \in I_c) = E(x'_{it}|Z_{it}, s_{it}=1 |g_i \in I_c)\beta + E(u_{it}|Z_{it}, s_{it}=1 |g_i \in I_c); i= 1,.,N_t; t= 1,.,T \quad (9)$$

Expression in (9) highlights two relevant features of the RCS model with sample selection. Firstly, that the sample counterpart of the conditional expectations of interest and determinant variables are not simple cohort-means of observed values but weighted means with conditional probabilities of selected values as weights. Secondly, that using (9) as an errors-in-variable estimating equation leads to inconsistent estimates unless $E(u_{it}| Z_{it}, s_{it}=1 |g_i \in I_c)$ is zero or time invariant. In the case that selection is time invariant FE estimators not only remove fixed effects but also eliminate selection biases. It can be noted that in the transit between individual and cohort data the emphasis goes from the probability of being observed, in the cross-section model, to the conditional probability of being observed given a specific cohort, in the RCS model.

A solution to achieve consistency is modeling $E(u_{it}| Z'_{it}, s_{it}=1 |g_i \in I_c)$. To cover the main characteristics of the panel data literature we must assume that u_{it} is a compound error with two components: individual effect and idiosyncratic error. As in Ridder and Moffitt (2007) the sample main equation we consider is a linear individual effects regression

$$Y_{it} = \beta_1 X_{it} + \delta Z_{i0} + f_i + \epsilon_{it} , \quad (10)$$

where f_i are individual effects; ϵ_{it} idiosyncratic errors; X_{it} are time-varying variables (tvc) and Z_{i0} time invariant variables (tic). Fixed effects are potentially correlated with X_{it}, Z_{i0}. Usually Z_{i0} is a dummy cohort-indicators matrix.

The selection equation is a time-varying selection mechanism, Semykina and Wooldridge (2010),

$$s_{it} = \gamma_t Z_{1it} + f_i + \varepsilon_{it} , \quad (11)$$

where s_{it} is a dichotomous variable that takes 1, 0 values (1 when individual i is observed, 0 otherwise); Z_{1it} is a matrix of determinants of the selection process. Z_{1it} relevant terms are time-varying variables but does not exclude time-invariant covariates. Y_{it} is only observed when s_{it} =1. Due to the time-varying assumption the fixed effects in (11) are unidentified in the cross-section but can be approximated, Semykina and Wooldridge (2010), by Mundlak's (1978) modeling procedure. Taking expectations in (11) for fixed t gives

$$E(s_i| g_i \in I_c) = 1*[Prob(s_i =1| g_i \in I_c)] + 0*[Prob(s_i =0| s_i =1, g_i \in I_c)] = Prob(s_i =1| g_i \in I_c). \quad (12)$$

This expression, as we will see in the next section, form the basis to use the selection equation as a relevant element to estimate a bias correction term for the main equation.

3. Identification and selection-bias correction term modeling.

As stated before cohort variables are used in a RCS model as matching instruments. To estimate the system of equations (10) and (11), or their equivalent cohort system, we need a set of identifying restrictions. Although we allow for two sources of selection biases we assume the only nonzero time-varying expectation arises from the idiosyncratic errors. Our approach is in the line of Gronau (1985) and encompasses Moscarini and Vella (2002). As in Gronau´s work there is a time-varying source of selectivity bias that comes from the idiosyncratic terms; however, we take into account tvc variables, different from time and non-monotonous with respect to time,

could play an important role in determining the selection process. As in Moscarini and Vella´s research there is a time-invariant source of selectivity bias that comes from individual effects and therefore can be eliminated through FE estimators.

Assumption 1.

1.a (Z_{1i}, s_i) are observables; (y_i, X_i) are observed when $s_i = 1$[1].

1.b (u_i, v_i) independent from Z_{1i} and $E[u_i| Z_{1i}] = E[v_i| Z_{1i}] = 0$.

1.c v_i is distributed as N (0,1).

1.d $E[u_i|v_i]= \rho v_i$.

1.e $E(\lambda_{ct}| \lambda_{ct})= \lambda_c$.

1.f $E(e_{ct}| \lambda_{ct})= E(\lambda_{ct}| \eta_{ct})=0$.

Assumption 1.d holds for instance when we assume the idiosyncratic errors of both equations are jointly bivariate normally distributed.

Under assumption 1.d, a linear projection of u_{it} onto v_{it} is

$$u_{it} = \rho v_{it} + \eta_{it} ,$$

where η_{it} is independent of v_{it}. The relevant bias correction term in equation (9) becomes

$$E(u_{it}|Z_{it}, s_{it}=1|g_i \in I_c\, s_{it})= E (\rho v_{it} + \eta_{it} |Z_{it}, s_{it}=1|g_i \in I_c) = \rho\, E (v_{it} |Z_{it}, s_{it}=1|g_i \in I_c) = \rho\, E (v_{it} | s_{it}=1|g_i \in I_c) . \quad (13)$$

If we denote $\alpha_{ct} = Prob(s_{it}=1 |g_i \in I_c)$, the time-varying conditional probability that an individual is observed given this individual is a member of a specific cohort, a standard statistic result we have reviewed above is that the expectation term is equal to the IMR with argument the normit of α_{ct}. A consistent estimation of this probability can be obtained from the selection equation. Substituting (13) in (9) gives

$$E(y_{it}|Z_{it}, s_{it}=1 |g_i \in I_c)= E(x'_{it}|Z_{it}, s_{it}=1 |g_i \in I_c)\, \beta_1 + \rho\, \lambda_{ct}(\alpha_{ct}); \; i= 1,.,N_t; \; t= 1,.,T \quad (14)$$

where $\lambda_{ct}(.)$ is the IMR. Then in the presence of selection bias due to a time-varying selection mechanism to achieve consistency in the estimation of (10) we have to augment the specification with an additional regressor, λ_{ct}. The value of this cohort-time regressor is fixed for all observed individuals in cohort c and time t.

[1] Remember that variables included in X could be always observed.

We have to note that a test for the presence of selection bias will involve testing the null hypothesis of $\rho=0$ in (14), that is Ho: $\rho=0$. As usual the test can be viewed as an omitted-variable test in (14).

The estimating augmented main equation is

$$Y_{it} = \beta_1 X_{it} + \delta Z_{i0} + \rho\, \lambda_{ct} + f_i + \epsilon_{it} \,. \qquad (15)$$

If we could observe λ_{ct} an IV estimation would give a consistent estimation of the parameters of the model. As λ_{ct} depends on unknown parameters this direct procedure is unviable. In the next section we will present a generalized method of moments corrected (GMMC) estimator. Equation (15) is in the line of the seminal contribution of G-L. The Gronau suggestion of correcting for selection bias in the cohort equation with an additional term equal to the IMR with argument the normit of the observed proportions of the individuals in each cohort (proportion of 1 in each cohort) implies that the consistent estimation of the Prob(s_{it}=1 $|g_i \in I_c$) can be obtained through a linear specification (a linear probability model) of the selection equation in which the time-varying selection mechanism only depends on cohort dummies, as we will see later on.

Deaton (1985) shows that an errors-in-variables pseudo panel model can be a good approximation to the population model. It implies that IV moments equation derived from (15) must be modified to account for the presence of measurement errors. This suggest a generalized method of moments corrected (GMMC) system. Formally, moments equation associated with (15) is

$$E[(Y_{it} - X'_{it}\beta_1 - Z'_{0i}\delta - \rho\lambda_{ct})h(Z_{0i}, Z_{2it})] = B\beta + b \qquad (16)$$

where Z_{2it} are time-varying instrumental variables; h(.) is a known function usually a set of time and cohort-time interactions although any other time-varying variable is not discarded; $\beta = (\beta_1' \; \delta' \; \rho)'$; B, b depend on the covariance matrix of the measurement errors. For known λ_{ct} moments equation in (16) has been extensively studied in the literature, Deaton (1985), Verbeek and Nijman (1992, 1993), Moffitt (1993), Verbeek (1996), Collado (1997), McKenzie (2004), Devereux (2003), Ridder and Moffitt (2007). Properties of moment estimators are well known in general, Hansen (1982).

We will now proceed to the discussion of the way of modeling the argument of the IMR bias correction term in the main equation.

In G-L the argument of the IMR is the normit of the proportion of 1 in each cohort. This term is cohort-time specific. G-L suggestion has been often used in the empirical labour literature, see for example Blundell et al. (1998). Our suggestion (MM) is to use as argument of the IMR the normit of a consistent estimation of the conditional probability that an individual is observed given a specific cohort membership. Our approach is equivalent to consider relevant the distinction between the observed conditional probability (real proportion of 1 in each cohort-

time) and a consistent estimation of this conditional probability that takes into account their observed determinants (consistent estimation obtained through a selection rule equation). Many arguments can be given to support the idea that improving the specification of the selection equation 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, 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.[2]

Under the assumption of a time-varying selection equation, G-L and MM can be expressed in formal terms as the result of an OLS estimation of two different individual data regressions for each cross-section.

G-L suggestion can be interpreted as a cross-section regression that in matrix expression is

$$S = Z_0\alpha + \varepsilon \,, \tag{17}$$

where $S\{s_i\}$ is a N_t column vector that contains a selection variable (it takes the value 1 when an individual is observed and 0 otherwise); Z_0 is a (N_t x C) dummy cohort indicators matrix; α is a (C x 1) parameter vector and ε a column vector of error terms; N_t is the cross-section sample size. An OLS estimator of α

$$\hat{\alpha} = (Z'_0 Z_0)^{-1} Z'_0 S = AS, \tag{18}$$

gives us a column (Cx1) vector of proportions of 1 in each cohort, $\hat{\alpha}\{a_c\}$. Matrix $A = (Z'_0 Z_0)^{-1} Z'_0$ is a cohort-means operator.[3] A trivial feature of equation (17) is that is fully determined ($R^2=1$) so their estimations are not subject to errors. In G-L the observed proportions a_c are "true" values of the conditional probability of being observed given a specific cohort membership.

MM regression is a reduced-form equation

$$S = Z_1\gamma + \varepsilon \tag{19}$$

[2] Needless to say that the equivalence between G-L and MM 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 time-invariant variables just to preserve the desired size.

[3] The same result can be obtained if we assume that selection is a purely random process. Given the result of the selection rule an application of operator Z_0 to S produces the cohort-aggregates vector that must be divided by the total number of individuals in each cohort to get the observed proportions.

where Z_1 is a matrix of selection process determinants. Z_1 relevant terms are time-varying variables but does not contain Z_0 .[4]Otherwise if Z_0 were a subset of Z_1 results from MM and G-L regressions coincide.

To get consistent estimators of the conditional probability of being observed given a specific cohort membership we premultiply by matrix A. Then

$$AS=AZ_1\gamma + A\varepsilon \tag{20}$$

An OLS estimator of γ is

$$\hat{\gamma} = [Z'_1A'A\,Z_1]^{-1}\,Z'_1A'A\,S= \gamma + [Z'_1A'A\,Z_1]^{-1}\,Z'_1A'A\,\varepsilon \tag{21}$$

An estimation of the required probabilities and its covariance matrix is

$$A\hat{S} = AZ_1(Z'_1A'AZ_1)^{-1}Z'_1A'AS \tag{22}$$

$$var(A\hat{S}) = AZ_1(Z'_1A'AZ_1)^{-1}Z'_1A'A\Omega A'AZ_1(Z'_1A'AZ_1)^{-1}Z'_1A' \tag{23}$$

A consistent estimation of the covariance matrix can be obtained by White method. Then, the additional bias correction regressor to be included in our main equation will be for all individuals in each cohort-time the IMR with argument equal to $\Phi^{-1}[(Z'_0Z_0)^{-1}Z'_0\hat{S}]$.

An alternative view of this procedure comes from the equivalence between IV estimation and estimation with aggregated data. An IV estimation of the initial equation using A as instruments matrix leads to the same results. A proof for two-stage least squares (linear projection of Z_1 onto Z_0) is straightforward. Our procedure respects the RCS spirit and estimate the relevant conditional probabilities from cohort-means data.

All regressions have heteroscedastic disturbances. This fact derives from the choice of a linear probability model approach to represent the selection rule. So far the assumption of a linear probability specification for the selection rule has been maintained in order to establish a simple comparison with Gronau's procedure. In the case we use a probit to model the selection rule we have to consider proportions estimation such as, for example, Greene (2003).

---

[4] To directly obtain a vector of consistent estimates of the conditional probability of being observed given cohort membership a trivial procedure can be used. We premultiply by $(Z'_0Z_0)^{-1}Z'_0$ the vector of predictions $\hat{S}$, i.e. we linearly project $\hat{S}$ onto Z_0,

$$(Z'_0Z_0)^{-1}Z'_0\hat{S} = (Z'_0Z_0)^{-1}Z'_0Z_1(Z'_1Z_1)^{-1}Z'_1S\ .$$

With covariance matrix

$$var\left[(Z'_0Z_0)^{-1}Z'_0\hat{S}\right] = (Z'_0Z_0)^{-1}Z'_0Z'_1(Z'_1Z_1)^{-1}Z'_1\Omega Z_1(Z'_1Z_1)^{-1}Z'_1Z_0(Z'_0Z_0)^{-1}.$$

A consistent estimation can be obtained by White method.

For each t moments equation associated with (20) is

$$E[(s_{it} - Z'_{1it}\gamma_t)A_t] = 0 \quad (24)$$

Up to now we have not considered the presence of cohort fixed-effects in the selection equation estimation just to preserve simplicity in the comparison. When biases arising from right-hand side variables and fixed-effects correlation in the selection equation are relevant we use Mundlak's (1978) modeling device and augment the right-hand side variables with time average of cohort values of the included variables, Semikyna and Wooldridge (2010). The advantage of this approach is that it conserves on degrees of freedom.

4. Pseudo panel data and selectivity bias: A GMMC approach.

In this section we outline a formal GMMC approach to estimate our system. We will see that our GMMC estimators are a type of two-step estimators. Firstly, we estimate the MM system in (24) and the estimated parameters will allow us to estimate the additional regressor in (16); secondly, we estimate the GMMC system in (16). As estimating (24) is straightforward in the rest of the section we will devote our attention to consider questions relative to the estimation of (16) and to provide the covariance matrix of the estimators and an upper bound for the covariance-corrected matrix of the estimators due to the presence of the estimated additional regressor.

In cohort form the set of moments equations (24) and (16) can be expressed as

$$E[s_{ct} - Z'_{1ct}\gamma_t] = 0; \text{ t=1, 2....T, c=1, 2....C,} \quad (25)$$

$$E[(\Delta Y_{ct} - \Delta X'_{ct}\beta_1 - \rho\Delta\lambda_{ct})\Delta W_{ct}] = B\beta + b \ , \quad (26)$$

where $\Delta W_{ct} = (\Delta X'_{ct},\ \Delta\lambda_{ct})'$. Equation (25) is a system of T cross-section linear regressions. For probit specifications of the selection rule (25) would have to be modified to accommodate to a system of proportions regressions. In equation (26) we have used first-differences of the synthetic panel, one of the alternatives suggested by Deaton (1985). Substituting $\widehat{\gamma_{ct}}$ in (26) we get

$$E\left[\left(\Delta Y_{ct} - \Delta X'_{ct}\beta_1 - \rho\Delta\widehat{\lambda_{ct}}\right)\Delta X_{ct}\right] = B\beta + b \ . \quad (27)$$

The GMMC estimator is

$$\hat{\beta} = \left[\sum_{c=1}^{C}(\Delta W'_c \Delta \mathrm{W}_c + B')D_c \sum_{c=1}^{C}(\Delta W'_c \Delta \mathrm{W}_c + B)\right]^{-1}$$

$\left[\sum_{c=1}^{C}(\Delta W'_c \Delta \mathrm{W}_c + B')D_c \sum_{c=1}^{C}(\Delta W'_c \Delta \mathrm{Y}_c - b)\right]$, (28)

where $\Delta \mathrm{W}_c = (\Delta \mathrm{W}_{c2}, \Delta \mathrm{W}_{c3}, \ldots\ldots \Delta \mathrm{W}_{cT})'$, $\Delta \mathrm{Y}_c = (\Delta \mathrm{Y}_{c2}, \Delta \mathrm{Y}_{c3}, \ldots\ldots \Delta \mathrm{Y}_{cT})'$. The optimal choice of $\mathrm{D_c}$, Hansen (1982), is any consistent estimator of the inverse of the covariance matrix of $\Delta W'_c \Delta \mathrm{W}_c$. The asymptotic distribution of the GMMC estimator, for B, b, $\Delta \mathrm{W}_c$ known, can be derived using standard assumptions and GMM theory. The covariance matrix of the GMMC can be found for example in Collado (1997).

Let

$$m_{1ct} = s_{ct} - Z'_{ct}\gamma_t \text{ , t= 1,.....T}$$
$$m_{2ct} = (\Delta Y_{ct} - \Delta W'_{ct}\beta)\Delta W_{ct}, \text{ c=1,2,...C, t=2,3,...T}$$

The sample averages are

$$\bar{m}_{1ct}(\gamma_t) = \frac{1}{C}\sum_{c=1}^{C}(s_{ct} - Z'_{ct}\gamma_t) \text{ , t= 1,.....T} \quad (29)$$

$$\bar{m}_{2ct}(\gamma_t, \beta) = \frac{1}{C}\sum_{c=1}^{C}((\Delta Y_{ct} - \beta\ \Delta W'_{ct})\Delta W_{ct}), \text{ c=1,2,...C, t=2,3,...T} \quad (30)$$

Let $m(\theta) = (\bar{m}_{1c2}, \bar{m}_{1c3} \ldots\ldots \bar{m}_{1cT}, \bar{m}_{2ct})'$; $\theta = (\gamma'_t, \beta')'$. The system moments equation can be written in stacked form as m=0. This system correspond to a two-step-GMMC estimation. We have to estimate in the first step T independent regressions and then construct the estimated values of IMR. In the second step we estimate a measurement errors corrected (T-1) synthetic cohorts regression “a la Deaton”.

To get a consistent estimator of the asymptotic variance of $\hat{\theta}$ we need the following jacobian terms,

$$\hat{\mathrm{G}}_{\beta c} = \nabla_\beta \bar{m}_{2ct}(\gamma_t, \beta) \quad (31)$$

$$\hat{\mathrm{G}}_{\gamma_t c} = \nabla_{\gamma_t} \bar{m}_{2ct}(\gamma_t, \beta), \text{ t=1,2.....T} \quad (32)$$

$$\hat{\mathrm{M}}_{ct} = \nabla_{\gamma_t} \bar{m}_{1ct}(\gamma_t), \text{ t=1,2.....T} \quad (33)$$

Let $\hat{\mathrm{M}}_c = \{\hat{\mathrm{M}}_{ct}\}$ a (TxT) diagonal matrix; $\hat{\mathrm{G}}_{\gamma c} = (\hat{\mathrm{G}}_{\gamma_1 c}, \ldots\ldots, \hat{\mathrm{G}}_{\gamma_T c})$ a (1xT) row vector; 0 a (Tx1) column vector of zeros. Then $\hat{G}$ is a (T+1) squared lower triangular matrix

$$\hat{G} = \begin{bmatrix} \hat{\mathrm{M}}_c & 0 \\ \hat{\mathrm{G}}_{\gamma c} & \hat{\mathrm{G}}_{\beta c} \end{bmatrix} \tag{34}$$

Let (a consistent estimation of)

$$\hat{\Omega} = m_\theta m_\theta{}' \tag{35}$$

A consistent estimator of the covariance matrix of $\hat{\theta}$ is

$$V_\theta = G^{-1}\hat{\Omega}(G^{-1})' \tag{36}$$

Our interest, however, is in the covariance matrix of $\hat{\beta}$, the main equation parameters estimators. Its theoretical derivation is complex and in our view of little help for empirical research. We will give instead, following Deaton (1985) and Newey and McFadden (1994), a convenient expression for an upper bound of the covariance matrix V_β. The formula is

$$V_\beta = [\mathrm{M_{WW}} - \Sigma]^{-1}\left[\Sigma_{\mathrm{WW}}\left(\sigma_\mu^2 + \sigma_{00} + \theta'\Sigma\theta - 2\sigma'\theta\right) + (\sigma - \Sigma\theta)(\sigma - \Sigma\theta)\ '\right][\mathrm{M_{WW}} - \Sigma]^{-1} + \Pi'\hat{\nabla}\Pi \tag{37}$$

In equation (37) the first additive term is the covariance matrix for a static pseudo panel data model (Deaton 1985:118). The second is the correction matrix required for using in the estimation of the pseudo panel data model an estimated regressor instead of the "true" regressor in the second-step of the two-step-GMMC estimation procedure. Newey and McFadden (1994) establishes that in general the estimated regressor causes a bias in the estimated covariance matrix, but the problem arises when the estimated regressor downward bias the estimated covariance matrix. They give a sufficient condition for the downward bias and outline the correction that has to be made for each cross-section through a weighted inner product. Let β be (lx1) and γ (kx1) vectors, to construct the inner product matrix we make a regression of a set of k variables (estimated coefficient of IMR in the second step of the GMMC procedure times the derivative of IMR (evaluated at argument value) times the derivative of the normit (evaluated at argument value) times the vector of regressors in the selection equation) on the complete set of regressors, l, in the main equation. This gives us a (kxl) matrix of estimated parameters. For all t, we stack all the (lxk) transpose matrices to form a (lxkT) matrix of stacked estimated coefficients. The weighting matrix is a (kTxkT) block-diagonal matrix whose main diagonal elements are the cross-section covariance matrices in the selection equation. Then for each cross-section, being $\hat{P}_t$ the (kxl) matrix of estimated coefficients, $\hat{V}_t$ the (kxk) cross-section selection equation covariance matrix in (17), the correction term is

$$\widehat{P'}_t \widehat{V}_t \widehat{P}_t \quad (38)$$

5.- Monte Carlo Simulations of the Testing Procedure

We run a Monte Carlo experiment to investigate the power of the Gronau and MM selection bias tests. First, 2000 individuals in 10 times period was simulated. The individuals were split in 10 equal cohorts in each time period. Thus, we follow the cohort in all time period and cohort dummies are used to keep track of individuals over time [see Vella and Verbeek (2005), Girma (2001), and Verbeek and Nijman (1993)]. Also, we simulate covariates and latent selection as,

$$X_{i(t),t} = Z_{i(t),t} + \omega_{i(t),t} \quad (37)$$

$$S^*_{i(t),t} = 1[r_{i(t),t} > 0] \quad (38)$$

In equation (37) $Z_{i(t),t}$ consists of 10 dummies of cohorts with identical probability in each time period; $\omega_{i(t),t}$ was generated at random from a normal distribution; and $r_{i(t),t}$ was generated at random from a uniform normal distribution N [0,1]. The main equation was generated as follows,

$$Y_{i(t),t} = X_{i(t),t} + \varphi_{i(t),t} \quad (39)$$

Now, we make several hypotheses about the selection mechanism,

a. Gronau Selection

$$\lambda_{i(t),t} = \Phi(n_{i(t),t}/N_{i(t),t}) / \Phi(n_{i(t),t}/N_{i(t),t}) \quad (40)$$

b. MM Selection

$$S_{i(t),t} = Z_{i(t),t} + \eta_{i(t),t} \; ; \; \lambda_{i(t),t} = \Phi(\Phi^{-1}[\text{Prob}(S_{i(t),t}=1|\text{ gi Ic})]) / \Phi(\Phi^{-1}[\text{Prob}(S_{i\,i(t),t}=1|\text{ gi Ic})]) \quad (41)$$

The main equation to estimate is,

$$Y_{i(t),t} = \beta' X_{i(t),t} + C_{i(t)} + \gamma' \lambda_{i(t),t} + \psi_{i(t),t} \quad (42)$$

We test the power of the test based on the null hypothesis that γ equals zero. We make 2,000 iterations, 10 cohorts and 10 time periods and discard a 10% of individuals at the initial time (t=0). The corresponding results are listed below:

Table 1: n_c=200, C=10, T=10.

	β	Power
γ=0	0,9819	
Gronau	1,0463	0,0860
MM	0,9585	0,0450

Table 1, shows the simulation results using the nominal size of 5% as the benchmark. The first file shows the mean of the β without including the inverse mills ratio in the regression. The power of reject the hypothesis of γ=0 is an 8% in Gronau and 4.5% in MM test.

Next we discuss the effects of reducing a 10% of the individuals in each period versus don't reduce the number of individuals in each time period,

Table 2: C=10, T=10

No Reduction			Reduction of 10%		
	β	Power		β	Power
γ=0	1,0028		γ=0	1,0683	
Gronau	0,3227	0,0480	Gronau	0,3332	0,0800
MM	0,9736	0,0430	MM + r_t	0,9375	0,0400

The power of the test without a reduction of individuals is around 5% in both tests. But when the pseudo panel experiences a reduction of a 10% of individuals the power of Gronau test is worse than its power without reduction. Yet the power of MM test is better in the former than in the latter case.

In order to discuss the size of the tests we consider the following selection model,

$$S^*_{i(t),t} = 1[r_{i(t),t} + \eta_{i(t),t} > 0] \quad (43)$$

$$Y_{i(t),t} = X_{i(t),t} + \varphi_{i(t),t} \quad (44)$$

$$\text{Corr}(\eta_{i(t),t}, \varphi_{i(t),t}) = \rho \quad (45)$$

We have made 2,000 iterations, 10 cohorts, t=10, and reduced a 10% of individuals in each period and have used a bivariate normal distribution to simulate $\eta_{i(t),t}$ and $\varphi_{i(t),t}$. In order to consider the size of the test we estimate the equation (42) and use the nominal size of 5% as the benchmark to reject the hypothesis of $\gamma \neq 0$. The results are in Table 3 below.

Table 3

ρ	Gronau	MM
0.5	0.046	0.031
0.6	0.040	0.026
0.7	0.032	0.024
0.8	0.034	0.024
0.9	0.034	0.020

Results show that both tests are significant around 5%. However, MM test performs better than Gronau test.

Finally, in order to consider the monotonicity with respect to time we include a time-varying variable in the selection process and discuss the sample selection bias using 10,000 iterations, 10 cohorts and a bivariate normal distribution to simulate $\eta_{i(t),t}$ and $\varphi_{i(t),t}$. and consider a correlation of the 0.9. The results are,

Table 4

Test/Significance	1%	5%	10%
Gronau	0.9876	0.9408	0.8841
MM	0.0005	0.0007	0.0010

Results in table 4 show a poor performance of the Gronau test compared to MM test. In all cases, MM detect the sample selection bias.

6. Empirical application of the test: Estimating the returns to education

The return to education has been discussed in deeply around the world. In particular the econometric estimation of the Mincer equation, in honor to Mincer (1962), let us estimate the return to an additional year of education. In Colombia, the returns are almost 15% in the last century, before in the nineties was around 8%. A few articles in the Colombian literature discuss the selection problem. In particular, in this period only Tenjo and Bernat (2002) made

corrections of the returns to education by selection bias in cross-sections. We run a Mincer equation and test the existence of selection bias.[5] The main and selection equations system is,

$$lwh_{i(t),t} = \alpha_{i(t)} + \beta_o' E_{i(t),t} + \beta_1' Exp_{i(t),t} + \beta_2' Exp^2_{i(t),t} + \rho E(\xi_{i(t),t} | s_{i(t),t}) + \mu_{i(t),t}\,;\ t = 1,....,T\ ;i=1,...,N \quad (46)$$

$$S_{i(t),t} = E_{i(t),t} + N_{i(t),t} + \eta_{i(t),t} \quad (47)$$

Where $lwh_{i(t),t}$ is a logarithm of the wages by hour. $E_{i(t),t}$ are years of education. $Exp_{i(t),t}$ is a potential experience ($Age_{(t),t} - E_{i(t),t} - 6$) and squared of potential experience, $Exp^2_{i(t),t}$. $\alpha_{i(t)}$ is non-observable individual heterogeneity and $\mu_{i(t),t}$ is the error in each period and individual. The term $\rho\ E(\xi_{i(t),t} \mid s_{i(t),t})$ implies the existence of selection biases in the wage equation due we observe only employment individuals. In order to considerer the selection process, we define the labor participation, $S_{i(t),t}$, as a dummy variable that take value of 1 when the individual participate in the labor market (work or unemployed) and 0 in other case and use years of education, $E_{i(t),t}$, and the number of individuals $N_{i(t),t}$, at home (as proxy of the change in the cost of the labour search) as co variables for the selection process.

In terms of the sign of the parameters we expect a positive sign for years of education and potential experience. However, due to life of cycle we expect a negative sign for squared potential experience.

In Colombia there is no panel survey statistics on household labor supply data. Our sample comes from the National Housing Survey (NHS) which consists of a time series of independent and representative cross-sections collected from 1984 to 2000 by the National Agency of Statistics (DANE). Since 2000, the DANE has collected information about the labor market through another mechanism called Continuous Housing Survey[6].

In each year, the modules of working individuals, personal characteristics, work force, and education were linked. The data for variables as schooling years, age, labor earnings, household size, and number of working hours, were obtained through this link. In this way, the observations are independent cross-sectional series where individuals are only available in each period. Since there are different individuals in each period, i range from 1 to N for each t. In this case, we define five cohorts with 16 and 44 years old. The variables for schooling years, age, labor earnings, number of working hours, married, and kind of occupation were obtained from this procedure. We have 85,540 individuals in the total sample consisting of 39,015 women and 46,525 men.

[5] Mora and Muro (2008) discuss the additional returns to diploma in Colombia using Pseudo Panel data.
[6] Because of this information before and after 2000 is not comparable.

Table 5. Number of individuals by Cohort

Year	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Total
1996	2048	3535	4161	3853	3547	17144
1997	2347	3805	4064	4094	3341	17651
1998	2691	3693	3959	3729	3384	17456
1999	2706	3558	3670	3668	3055	16657
2000	3014	3425	3590	3611	2992	16632

Source: Data from DANE-ECH.

In table 5 we have more than 2000 individuals by Cohort and the average individuals in Cohort 1 (Young people) are 2561 and the average individuals in Cohort 5 (Old people) are 3264 individuals.

Table 6. Mincer equation in Colombia (1996-2000).

Variable/Method	Pool	Deaton	Gronau	MM
Years of Education	0.175***	0.3778488***	0.138194***	0.1482052***
	(0.006)	(7.41e-06)	(4.71e-06)	(0.011575)
Potential Experience	0.031***	0.3588496**	0.0091198	0.0382054***
	(0.039)	(0.0000197)	(9.12e-06)	(0.0009938)
Potential Experience^2	-0.00003	-0.0073193***	0.0004797***	-0.0002055**
	(0.001)	(8.84e-07)	(2.97e-07)	(0.0000207)
Inverse Mills Ratio	3.009***		5.72532***	4.717903***
	(0.141987)		(0.0001227)	(0.1133548)

Note: Standard error in parenthesis; * p<0.05, ** p<0.01, *** p<0.001

Table 6 above show the returns to education in the period 1996 to 2000 using the standard Mincer equation. In pool regression we estimate a Heckman regression using the participation in the labor market as a selection variable and years of education and number of individuals in the household as covariates of the selection process.[7] Pooling show a 17.5 percent of the return to additional year of education.

In the second column we estimate a pseudo panel Mincer equation using Deaton (1985) method. In particular, we correct the measurement errors but we don't correct for the selection bias. The result shows an overestimate in the return of education and experience.

[7] In all cross sections regressions of the Mincer Equation the Inverse Mills was positive and statistically significant.

Third column, show the results of the Mincer equation using Gronau method. In particular, we compute the Mills Ratio using proportions of the cohort in each time period and include the Mills ratio in the main equation; The corrections of the measurement errors as in Deaton (1986). Results show a 13.8 percent of the return to additional year of education. However, the sign of the potential experience is different from the theory; that is, a negative sign. This results, is due to the inconsistence of the IMRG argument in the principal equation.

Finally, the table 6 shows the estimation of the Mincer equation using our method (MM). In each time period we estimate the selection process as the regression using the mean by cohorts of the participating in the labor market over years of education and the number of individuals in the household. Following, we collect the covariance matrix using the MM procedure in order to analize the selection bias. We use the deviation of the individual data from the cohort in order to correct measurement error as in Deaton (1985). Our results show a 14.8 percent of the return to additional year of education and 3.7 percent of the additional year of potential experience. All signs of the co-variables as in the theory and all co-variables show a statistical significance. Our method shows the returns comparable with other results for the period for Colombia [Prada (2006), Hernandez (2010)].

In order to consider the inconsistency of the standard error due the two steps GMM we make a correction of the covariance matrix in the spirit of the Newey – McFadden (1994). That is, in each time period we compute Phi Matrix as the result of the multivariate regression between $\hat{\alpha}_c \lambda_c(z_c \hat{\gamma}_c) z_c$ on S_c, exp_c, exp^2_c, $InvMills_c$. Finally, Phi Matrix was pre and post multiply by the covariance matrix in the first step (First step include $E_{i(t),t}$ and $N_{i(t),t}$ S co variables of the selection process).

7. Conclusions

In this article, we discussed a simple testing procedure for sample selection bias in pseudo panels. We described a pseudo panel model in which, under convenient expansion of the original specification with a selectivity bias correction term, the method allows to test for selection bias. We showed that the proposed selection bias correction term is proportional to Mills inverse ratio with an argument equal to the "normit" of a consistent estimation of the conditional probability of an individual is observed given cohort membership.

The test can be considered a cohort counterpart of Heckman's selectivity bias test for the individual case and, to some extent, generalizes previous existing results in the empirical labor literature.

In particular we propose an two-step-GMMC estimation. This procedure implies in order to achieve consistency the estimation in the first step T independent regressions and then construct the estimated values of IMR. In the second step we estimate a measurement errors corrected (T-1) synthetic cohorts regression "a la Deaton".

We discussed the power and size of the proposed test using Monte Carlo simulations. We made 2,000 iterations, 10 cohorts, t=10 and n_c = 200. Our simulations show that a comparison between two alternative tests, Gronau (1974) and ours, gives an 8% in Gronau and 4.5% in MM test. In the case of analysing the reduction of a 10% of individuals in each time period the power of Gronau's is worse than without reduction and MM's is better in the former than in the latter. Additionally, we used a bivariate normal distribution to simulate $\eta_{i(t),t}$ and $\varphi_{i(t),t}$ in order to consider the existence of selection bias in the main equation. Our results show that both tests are significant around 5%. However, MM test perform better to Gronau test. Finally, our results show a poor performance of the Gronau test compare to MM test when we included a time-varying non-monotonous with respect to time variable in the selection process.

Finally, we applied the proposed test and associated estimation to an empirical example. To analyse the Mincer returns to education for the Colombian labour market using Gronau and MM method as a correction of the selection bias. Our results show the existence of selection bias and a clear relevance of the test to obtain consistent estimators. In order to consider the inconsistency of the standard error due the two steps GMMC we make a correction of the covariance matrix computing in each time period the Phi Matrix as the result of the multivariate regression. Our results shows a 15 percent of the return to additional year of education in the Colombian labor market for the period 1996 and 2000.

8. References

Blundell, R., A. Duncan, and 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., and M. Rochina-Barrachina (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.

Greene, W.H (2003). *Econometric Analysis*. Pearson Education.

Heckman, J (1979), "Sample selection bias as a Specification Error", *Econometrica* **47**: 153-161.

Jensen, P., Rosholm, and M. Verner (2002), "A Comparison of Different Estimators for Panel Data Sample Selection Models", University of AARHUS, W.P. No. 2002-1.

Hernandez, G (2010), ¿Cuán rentable es la educación superior en Colombia? Lecturas de Economía **73**: 181-214.

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.

Meijer, E., and T. Wansbeek (2007), “The Sample Selection Model from a Method of Moments Perspective”, *Econometrics Reviews* **26**(1): 25-51.

Moffitt, R (1991), “Identification and estimation of Dynamic Models with a Time Series of Repeated Cross-Sections”, Brown University, Providence RI, mimeo.

Moffitt, R (1993), “Identification and estimation of Dynamic Models with a Time Series of Repeated Cross-Sections”, *Journal of Econometrics* **59**: 99-123.

Mora, J.J., and J. Muro (2008), "Sheepskin effects by cohorts in Colombia," *International Journal of Manpower* **29**(2): 111-121.

Moscarini, G., and F. Vella (2002), “Aggregate Worker Reallocation and Occupational Mobility in the U.S.:1971-2000”, IFS Working Papers, W02/18.

Mundlak, Y. (1978), “On the Pooling of Time Series and Cross-section Data”, *Econometrica* **46**: 69-85.

Prada, C.F (2006), “Is the Decision of Study in Colombia Profitable?”, *Ensayos sobre Política Económica* **51**: 226-323.

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 J.J. Heckman and E.E. Leamer (eds.) *Handbook of Econometrics* ***6***: 5469-5547.

Semykina, A., and J.M. Wooldridge (2010), ”Estimating Panel Data Models in the Presence of Endogeneity and Selection”, *Journal of Econometrics* **157**: 375-380.

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., and M. Verbeek (1999), “Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias”, *Journal of Econometrics* **90**: 239-263

Vella, F., and M. Verbeek (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.

Wooldridge, J.W (2002), *Econometric Analysis of Cross Section and Panel Data*. The MIT press.