Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2012-26 | May 24, 2012 | http://www.economics-ejournal.org/economics/discussionpapers/2012-26

Consistent Estimation of Pseudo Panels in the Presence of Selection Bias

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

Please cite the corresponding journal article:
http://dx.doi.org/10.5018/economics-ejournal.ja.2014-43

Abstract In the presence of selection bias, traditional estimators of pseudo panel data are inconsistent. In this paper, the authors derive the conditions under which consistence is achieved in pseudo-panel estimation and propose a simple test of selection bias. Specifically, they propose a Wald test for the null hypothesis that there is no selection bias. Under rejection of the null hypothesis, the authors can consistently estimate pseudo-panel parameters. They use cross sections and pseudo-panel regressions to test for selection bias and estimate the returns to education in Colombia. The authors corroborate the existence of selection bias and find that returns to education are around twenty percent.

JEL C23 C52

Keywords Repeated cross-section models, selectivity bias testing, human capital

Correspondence Jhon James Mora Rodriguez, Universidad Icesi, Colombia, e-mail: jjmora@icesi.edu.co

1. Introduction

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 Verner (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.

Moscarini and Vella (2002) discuss the sample selection in the context of the pseudo panel in a mobility model, in which, mobility and labour market participation equation errors are correlated. However, Moscarini's and Vella's (2002) do not discuss the presence of measurement errors in variables or the existence of measurement errors in the selection variable and the effects over consistence of the estimators. Off course, if we observe different individuals every period, we will obtain inconsistent estimators unless a set of assumptions on the selection process is established.

This paper shows 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.

In empirical applications of the test, we use Colombian labour data to estimate the returns of education between 1996 and 2000. Cross sections and pseudo panel regressions of the returns are estimate and test selection bias is made. We find that the returns of education are around a twenty percent and corroborate the existence of selection bias in the returns of education in Colombia.

The paper is structured as follows: Section two discusses the selectivity bias in a pseudo panel data. Section three present selectivity bias correction term for pseudo panel models. Section fourth discusses sample selection bias in the returns to education of Colombia. Finally, section five present conclusions.

2. General Framework

Let $i_{(t)} = 1, \ldots, N_t$. $t = 1,\ldots,T$. That is, each individual in each time period is different. In this way, we have a classical pseudo panel data model,

$$y_{i(t),t} = x'_{i(t),t}\beta + \alpha_{i(t)} + u_{i(t),t} \quad ; \tag{1}$$

We denote $y_{i(t),t}$ as an interest variable in repeated cross section model with measurement error.[1] $\alpha_{i(t)}$ are individual effects in t; $u_{i(t),t}$ 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 (1) 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 (1). 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 (1) when we have repeated and independent cross-sections data even in the case of correlation between individual effects and explanatory variables. Moffitt (1993) and Ridder and Moffitt (2007) 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

$$\alpha_{i(t)} = \sum_{c=1}^{C} d'_c \alpha_c + \tau_{i(t)} \ , \tag{2}$$

Where d_c is equal to 1 if individual i belong to cohort c and 0 otherwise. Substituting (2) in (1) we obtain

$$y_{i(t),t} = x'_{i(t),t}\beta + \sum_{c=1}^{C} d'_c \alpha_c + \tau_{i(t)} + \mu_{i(t),t} \ ; \quad t = 1,\ldots,T. \tag{3}$$

In equation (3) 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 $x_{i(t),t}$,. Consistency conditions for the estimator 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).

Now, we know that under the presence of sample selection bias the estimators are inconsistent (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

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

appropriate way. In particular, unobservable effects and selectivity bias could be removed through differencing, but this method is unfeasible in pseudo panels.

Now, consider that the presence of sample selection bias, so that {yi(t),t , xi(t),t } are only observed when $s_{i(t),t}$ equals 1. Then the IV estimator is,

$$\hat{\beta}_{IV} = \left[\left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_c) \right) \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_c)' \right) \right]^{-1} \times \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{x}_{ct} - \bar{x}_c) \right) \left(\sum_{c=1}^{C} \sum_{t=1}^{T} s_{i(t),t} (\bar{y}_{ct} - \bar{y}_c) \right) \tag{4}$$

We need two assumptions to assure efficient estimators,

Assumption 1: For i(t)=1,…,n; t=1,..,T. The correlations between idiosyncratic errors and selection are zero. That is,

$$\operatorname*{plim}_{N_c \to \infty} \frac{1}{NT} \sum_{i(t)=1}^{N} \sum_{t=1}^{T} s_{i(t),t} \mu_{i(t),t} = 0 \tag{5}$$

Where N_c is the number of individuals in each cohort. So,

$$\lim_{N_c \to \infty} \frac{1}{NT} \left[E \left(\sum_{i(t)=1}^{N} \sum_{t=1}^{T} s_{i(t),t} \mu_{i(t),t} \right) \right] = \lim_{N_c \to \infty} \frac{1}{NT} \left[\sum_{i(t)=1}^{N} \sum_{t=1}^{T} E \left(s_{i(t),t} \mu_{i(t),t} \right) \right] = 0 \tag{6}$$

Assumption 2: For i(t)=1,…,n; t=1,..,T. The correlations between idiosyncratic errors and individual effects are zero. That is,

$$\operatorname*{plim}_{N_c \to \infty} \frac{1}{NT} \sum_{i(t)=1}^{N} \sum_{t=1}^{T} s_{i(t),t} \alpha_{i(t),t} = 0 \tag{7}$$

Consequently,

$$\lim_{N_c \to \infty} \frac{1}{NT} \left[E \left(\sum_{i(t)=1}^{N} \sum_{t=1}^{T} s_{i(t),t} \tau_{i(t),t} \right) \right] = \lim_{N_c \to \infty} \frac{1}{NT} \left[\sum_{i(t)=1}^{N} \sum_{t=1}^{T} E \left(s_{i(t),t} \tau_{i(t),t} \right) \right] = 0 \tag{8}$$

It is worth noting that Assumption 2 holds true because of the fact that the deviation of heterogeneity with respect to the cohort is independent from the selection process itself. However, Assumption 1 is more disputable if the individuals are not selected at random.

Assumption 3: For i(t)=1,…,n; t=1,..,T. Under Assumption 1, 2 the estimator $\hat{\beta}_{IV}$ is consistent for fixed T and $N_c \to \infty$. We can observe that,

$$\begin{aligned} p\lim \hat{\beta}_{IV} = \beta + p\lim\left[\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)\right)\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)'\right)\right]^{-1} \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)'\right) \\ \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}\tau_{i(t),t}\right) + p\lim\left[\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)\right)\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)'\right)\right]^{-1} \\ \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}(\bar{x}_{ct}-\bar{x}_c)'\right) \times p\lim\left(\frac{1}{NT}\sum_{i(t)=1}^{N}\sum_{t=1}^{T} s_{i(t),t}\mu_{i(t),t}\right) \end{aligned} \tag{9}$$

So,

$$p\lim \hat{\beta}_{IV} = \beta \tag{10}$$

3. Testing for Selection bias in Pseudo Panel Data.

Heckman (1979) show a classical correction in longitudinal data and, Wooldridge (1995), in the others, extends to panel Data. Gronau (1974) and Lewis (1974) present the discussion in grouped data, but "they do not investigate the statistical properties of the method or develop the micro version of the estimator" (Heckman 1979, 156). Following Heckman (1979), the regression for the subsample of available data in pseudo panel data will be as follows:

$$\begin{aligned} &E(y_{i(t),t} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c) = \\ &E(x'_{i(t),t}\beta + \alpha_{i(t)} + \mu_{i(t),t} \mid\ x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c) = \\ &E(x'_{i(t),t}\beta \mid x_{i(t),t}, s_{i(t),t}=1, g_{i(t)} \in I_c) + E(\alpha_{i(t)} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c) + E(\mu_{i(t),t} \mid x_{i(t),t}, s_{i(t),t}=1, g_{i(t)} \in I_c) \end{aligned} \tag{11}$$

In equation (11) $g_{i(t)} \in I_c$ 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 (11) implies the use of the sample mean of the variables in the respective cohorts. By Assumption 2 the second term, $E(\alpha_{i(t)} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c$, 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 are an incidental truncation.

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 behaviour of a cohort based on the behaviour of different individuals in the cohort.

Now, suppose the instruments for the cohort is always observed unlike $\{y_{i(t),t}\ ,\ x_{i(t),t}\}$, which are observed only when $s_{i(t),t}$ equals 1. And define a latent variable $s^*_{i(t),t}$ as,

$$s^*_{i(t),t} = r'_{i(t),t}\,\beta + \eta_{i(t)} + \xi_{i(t),t} \quad , \qquad i(t)=1,\ldots,n;\ t=1,..,T \tag{12}$$

Where ri(t),t is a set of instruments, included cohort instruments, ηi(t) represents non-observable individual heterogeneity and ξi(t),t is the error term. Then, the selection indicator will be as,

$$s_{i(t),t} = 1\,[s_{i(t),t}{}^* > 0] = 1[r'_{i(t),t}\,\beta + \eta_{i(t)} + \xi_{i(t),t} > 0] \tag{13}$$

In equation (13) 1[•] is the indicator function. Of course, as equation 2 above, $\eta_{i(t)}$ is result of a cohort effect plus individual deviation. Following the work of Heckman (1979) and Wooldridge (1995), let us assume that $\{\mu_{i(t),t}\ ,\ \xi_{i(t),t}\}$ is independent from $\{\alpha_{i(t),t}\ ,\ \eta_{i(t),t}\}$. Thus, if $E(\mu_{i(t),t} \mid \xi_{i(t),t})$ is linear, then

Assumption 4: For i(t)=1,…,n; t=1,..,T.If Assumption 3 is hold, under incidental truncation, the expression for selection is equivalent to,

$$E(\mu_{i(t),t} \mid x_{i(t),t},\ s_{i(t),t}=1, g_{i(t)} \in I_c) = \rho\, E(\xi_{i(t),t} \mid r_{i(t),t}\ ,\ s_{i(t),t}) \tag{14}$$

Observe, if Assumption 4 is hold, then the main equation imply that,

$$y_{i(t),t} = x'_{i(t),t}\beta + \alpha_{i(t)} + E(\xi_{i(t),t} \mid r_{i(t),t}, s_{i(t),t})\rho + \psi_{i(t),t} \quad ;\ E(\psi_{i(t),t} \mid \alpha_{i(t)}, x_{i(t),t}, s_{i(t),t}) = 0 \tag{15}$$

Following equation (14) and (15), and Assumption 4, we define:

$$E(\xi_{i(t),t} \mid r_{i(t),t}\ ,\ s_{i(t),t}=1) = \lambda_{i(t)}\,(r'_{i(t),t}\,\beta + \eta_{i(t)}) \tag{16}$$

In equation (16) above, $\lambda_{i(t)}$ 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.

It follows that if we know $E(\xi_{i(t),t} \mid s_{i(t),t})$ then a contrast about the existence of selection biases will involve contrasting the hypothesis of a lack of significance of ρ in (14), that is Ho: ρ=0. 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.[2]

In this way the methodology for the selectivity bias under null hypothesis 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.

4. Empirical Application of the Test: The Returns of 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 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 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.[3] The main equation is,

$$lwh_{i(t),t} = \beta' x_{i(t),t} + \eta_{i(t)} + \rho\, E(\xi_{i(t),t} \mid s_{i(t),t}) + \mu_{i(t),t} \quad ; \quad t = 1,\ldots,T \; ; \; i=1,\ldots,N \qquad (17)$$

Where $lwh_{i(t),t}$ is a logarithm of the wages by hour. $X_{i(t),t}$ are years of education, S, potential experience (years – S – 6) and squared of potential experience, whereas $\eta_{i(t)}$ 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 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. Because of this, information before and after 2000 is not comparable. 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, wealth, sector and married were obtained through this link. In this way, the observations are independent cross-sectional series where N 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

[2] In this paper we don't discuss efficient properties of estimators. To assure efficient estimators we can use Murphy-Topel(1985) corrections.

[3] Mora and Muro (2008) discuss the additional returns to diploma in Colombia using Pseudo Panel data.

44 years old. The variables for schooling years, age, labor earnings, number of working hours, wealth, married, and kind of occupation were obtained from this correlation. The results of Mincer equations are,

Table 1. Mincer equation in Colombia (1996-2000).

	Miencer 2000 b/se	Mincer 1999 b/se	Mincer1998 b/se	Mincer 1997 b/se	Mincer 1996 b/se	Selection b/se	Pseudo Panel b/se
S	.1412541***	.1383655***	.1361496***	.1334026***	.1310991***	.1235483***	.1966922***
	(.0020854)	(.0021267)	(.0019503)	(.0018447)	(.0018007)	(.0082333)	(.0025296)
Exp	.0321968***	.031183***	.0274739***	.0274661***	.0314444***		.0313174***
	(.0030359)	(.0031636)	(.0028892)	(.0028898)	(.0029092)		(.0023519)
Exp2	-.0002252*	-.0002392*	-.0001819*	-.0001407	-.0002601**		-.0002522***
	(.0000914)	(.0000984)	(.0000924)	(.0000952)	(.0000972)		(.0000495)
Secagri	.0463419	.0172519	.0882515	.2379514***	.1309352*		.0649296*
	(.0636033)	(.0514475)	(.0677446)	(.0709436)	(.0577078)		(.0264414)
Secmin	.0222697**	.0192572	.0289787**	.0200611**	.0300206***		.0216815***
	(.0069641)	(.0115905)	(.0088756)	(.006499)	(.0059085)		(.0031611)
Secman	.0162709	.0566842**	.0612738***	.0673478***	.1141031***		.0799379***
	(.0169497)	(.0174129)	(.0161236)	(.0152298)	(.0154673)		(.0071878)
Secelec	.3265506***	.2397046***	.3640251***	.371691***	.377353***		.3422365***
	(.0814834)	(.0715125)	(.0595921)	(.0616362)	(.0513145)		(.028072)
Seccons	.0162995	.1988307***	.2033678***	.2132793***	.2865993***		.1893083***
	(.0373177)	(.0320874)	(.0275541)	(.0262455)	(.0243556)		(.0118297)
Seccomer	-.1030272***	-.0776031***	-.0370049*	-.0210478	.0201626		.0143986
	(.0171044)	(.0181467)	(.0162366)	(.0156571)	(.0160807)		(.007411)
Sectrans	-.016381	.0647884*	.1103051***	.1490831***	.1991585***		.1211066***
	(.0276592)	(.029188)	(.0288187)	(.027028)	(.0271018)		(.0122834)
Secbanca	.127423***	.1304969***	.1246852***	.1608867***	.2117894***		.1393315***
	(.0243519)	(.0219294)	(.0209863)	(.0208988)	(.0203362)		(.0093515)
Wealth						-.4307426***	
						(.0336628)	
Houshold						.0441722*	
						(.0175185)	
Married							.148486***
							(.0054027)
IMR							-1.158224***
							(.315368)
Constant	5.476374***	5.509942***	5.421515***	5.274845***	5.00781***	-.9673858***	5.383606***
	(.0368708)	(.0371642)	(.0339355)	(.0315026)	(.0314719)	(.0846548)	(.062225)
Year Effects	No	No	No	No	No	No	Yes
Adj. R-Squ~e	.4361482	.3849463	.3990234	.389036	.3844045		.5000716
Number	8657	9120	10455	11124	11073	85540	50429
Log-Likel						240.0372	
Number	-7633.228	-8444.263	-9471.326	-9944.135	-9849.921		-43083.93
F	483.7454	441.3975	519.3292	523.0436	540.1585		835.175
J-Overident							0

Table 1 above show the cross sections returns to education for 1996 to 2000. The average returns are approximately a 13 percent. The selection column shows the participation and we use a IV-Probit with the five cohorts as instrument. In this estimation, $Wealth_{i(t),t}$ is a dummy for wealth, and $Houshold_{i(t),t}$ is a dummy for household size. We have 85,540 individuals in the total sample consisting of 39,015 women and 46,525 men. With regard to the participation model, the findings show that the participation in the labor market increases as the number of schooling years increases. The wealth results in a decrease of their participation in the labor market. The results for IV-Probit for selection show that all coefficients are statistically significant (wealth, number of individuals in the home, and married).

Finally, in the last column we estimate a pseudo panel returns to education in Colombia. The results show 19 percent of the return in this period. The results also show the existence of selection bias in the Mincer equation. Also we control of industries and incorporate dummys of the economic sectors such as agricultural, minery, electricity, manufacturing, building, trade, transports, and financial services.

5. Conclusions

Sample selection bias is common in econometric estimation. Despite the continuous generalization of panel data surveys most developing countries still collect microeconomic information on economic agents' behaviour by means of repeated independent and representative cross-sections. In this article, we show 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 with null hypothesis of ρ equal to cero 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 individuals in each cohort. The test can be considered a cohort counterpart of Heckman's selectivity bias test for the individual case and generalizes to some extent previous existing results in the empirical labour literature.

Finally, we apply the developed text in the context of the Mincer returns of education in Colombia labour market. Our result shows the existence of selection bias and it´s clear the relevant of the test to obtain consistent estimators.

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.

Mincer, Jacob (1962) "On-the-Job Training: Costs, Returns and Some Implications" Journal of Political Economy, 70(5) Part 2, S50-S79.

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

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

Murphy, K.M. and R.H. Topel (1985), "Estimation and Inference in Two-step Econometric Models", Journal of Business and Economic Statistics, 3, pp. 88-97.

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 Recent Advances in Econometrics Methods. Edited by J. Heckman and Leamer E.

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.

Verbeek, Marno & Nijman, Theo, 1993. "Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections," Journal of Econometrics, Elsevier, vol. 59(1-2), pages 125-136,

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.