Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada Statistique Canada

Canada

MODEL ESTIMATION FOR SELECTIVITY OF ATTRITION, EMPLOYMENT AND WAGES USING THE SURVEY OF LABOUR AND INCOME DYNAMICS

Brahim Boudarbat[1] and Lee Grenon[2]

ABSTRACT

To understand the selection biases in model estimation when using longitudinal survey panel microdata, we consider a structural model composed of three equations for non-attrition/response, employment and wages. The three equations are freely correlated. The model is estimated using microdata from 22,990 individuals who provided sufficient information in the first wave of the second panel of the SLID (i.e. in 1996). Results provide evidence for non randomness of attrition behaviour. Attritors and non-respondents likely are less attached to employment and come from low-income population. We find small positive, though significant, correlation between non-attrition and employment. In addition, wage equation estimates are generally overestimated when selection from non-attrition and employment is ignored. It seems that observed wages are on average higher than wages that would be observed if all the respondents initially selected remained in the sample.

KEY WORDS: Model Estimation; Attrition; Employment; Wages

1. INTRODUCTION

This study examines the selectivity for attrition within model estimation for a specific subsample of longitudinal respondents. We are concerned with longitudinal respondents who participated in the labour interview and were 16 to 64 years-old in the first year of panel 2 of the survey of labour and income dynamics (SLID), i.e. in 1996. This subsample provided a minimum of information necessary for estimation with the model we use. A significant proportion of this subsample become out of scope or non-respondents by the end of the panel (i.e., in 2001). Respondents become out-of-scope when they migrate away from the Canadian provinces, are institutionalized, or are deceased. The out-of-scope respondents are outside of the target population for SLID, and are not eligible to participate for the reference year. The other component of attrition is non-response which includes persons who can not be located or contacted, and those who completely refuse to participate. The non-respondents are potentially still in the target population of the survey, but are no longer participating.

As sampled individuals exit from the sample, the data set becomes less representative of the population from which the longitudinal sample was drawn if the attrition is non-random. Some American and European econometric studies have analyzed the effect of attrition within longitudinal data on model estimation (see for instance the Special Issue “Attrition in Longitudinal Surveys,” of *The Journal of Human Resources*, Spring, 1998, Vol. 33, No. 2). This literature generally indicates evidence that the labour market behaviour of attritors and participants is different, although ignoring the selection bias has a minimal or negligible impact on estimation. In this study, we aim at verifying whether the above consistent result regarding the effect of attrition on estimation applies to the SLID. For this purpose, we consider a structural model composed of three freely related equations for non-attrition/response, employment and wages. The relationship between these equations arises from the fact that the employment status is observed only for respondents to the labour interview, and that wages are observed only for the respondents who are

[1] Brahim Boudarbat, School of Industrial Relations, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montreal, QC, Canada H3C 3J7. Email: brahim.boudarbat@umontreal.ca

[2] Lee Grenon, Statistics Canada, British Columbia Inter-university Research Data Centre, Walter Koerner Library, room 202, 1958 Main Mall, Vancouver, BC, Canada V6T 1Z2. Email: bcirdc@interchange.ubc.ca

The authors would like to thank Statistics Canada for providing access to the SLID data used in this study. Brahim Boudarbat also gratefully acknowledges financial assistance from SSHRCC INE Grant "Globalization, technological revolutions and education,” while he was a post-doc at UBC.

employed. Thus, the model allows testing for the selectivity from attrition in both employment and wages equations. Our model is developed in Section 3. Then, we evaluate the model using microdata from 22,990 working-age respondents in our subsample. The latter is described in Section 2 and structural parameter estimates are presented in Section 4. Section 5 offers a short summary with concluding remarks.

2. DATA

The analysis here includes only longitudinal respondents who provided information in the labour interview to ensure that a minimum of information is available for comparing the characteristics of respondents who exit the sample in subsequent years to those of respondents who are still in scope and providing information at the end of the panel period. Over one-fifth of the longitudinal sample who were in-scope and responded to the first wave of the labour interview were out of scope or non-respondents to the labour interview by the last reference year of the panel.[3] As with all Statistics Canada surveys, SLID produces survey weights using post-stratification methodologies to adjust weights for attrition and ensure reliable data quality. The longitudinal person weight for the 1996 reference year was used in the model estimation (Levesque and Franklin, 2000).

The characteristics of attritors/non-respondents and respondents differ and suggest that attrition may not be random within the SLID longitudinal sample. Respondents who become attritors or non-respondents in the subsequent year are on average younger, to not be married, and to be immigrants than non-attritors. In addition, respondents who are living in an urban area or who moved during the reference year (t) are more likely to become attritors in year (t+1). Moreover, attritors are more likely to live in urban areas with larger populations in their residential area than non-attritors. Attrition is also higher when respondents live without a spouse or common-law partner. It follows that on average, respondents who would later become attritors have lower wages and salaries and lower total household income.

3. ECONOMETRIC SPECIFICATION

The sample used in the estimation of the model developed below, is of 22,990 longitudinal in-scope respondents who provided sufficient information in the labour interview and were age 16 to 64 years in 1996. By the end of 2001, 7,381 individuals or 32% of these respondents became attritors or non-respondants at least once. To ease the estimation of our model, we consider attrition as absorbing state, so respondents, who became out of scope or did not respond to the labour force interview during a wave, are considered as attritors for the subsequent waves. We ignore the fact some attritors were converted to labour interview respondents in the following waves. [4]

The model focuses on the possible correlation between three variables: wages, employment status and respondent status (attrition). Wages are only observed for employed respondents, and the employment status is only observed for respondents who provide information on their employment during the reference year. Therefore, data on employment status is censored (missing) for non-respondents to the labour interview, and wages are censored for non-respondents and for respondents who are not employed. If this two-level censorship is not random, results based on observed data are subject to selection bias. In order to evaluate this potential selection bias when estimating employment status and wages, we propose the following model. The two selection sources are depicted by the reduced-form equations (1) and (2) below:

3.1 Non-attrition and response criterion

$$a_{it}^{*} = Z_{i(t-1)}\theta + \varepsilon_{1it}\ ,\ i=1,\ldots,n;\ t=2,\ldots,T_i \qquad (1)$$

---

[3] Data quality for SLID can be assessed through a variety of measures of non-response and estimates of variance (Michaud and Webber, 1995). This study does not measure non-response within the sample, rather the analyses here considers the selectivity of attrition within a structural model for a specific subsample of respondents.

[4] The return to the sample may also be non-random, which may counterbalance or intensify the effect of attrition.

Where i indexes for individuals and t indexes for time periods (i.e. waves of the survey). Individual i is a respondent in period t ($a_{it} = 1$) if $a_{it}^* \geq 0$, and is an attritor in period t ($a_{it} = 0$) otherwise. Because of the assumption that attrition is an absorbing state, if $a_{it} = 0$, then $a_{it'} = 0$ for any $t' > t$. Since information for the current period is not available for individuals who exit from the sample or do not respond, we use lagged variables. The initial period of analysis of attrition is the second wave since all individuals responded in the first wave.

3.2 Employment criterion

$$e_{it}^* = X_{it}\alpha + \varepsilon_{2it}, \; i = 1,...,n, \; t = 1,...,T_i \tag{2}$$

Respondent i is employed ($e_{it} = 1$) if $e_{it}^* \geq 0$, and is not employed ($e_{it} = 0$) if $e_{it}^* < 0$. A person is not employed if she is either unemployed or not in the labour force. Z_{it} and X_{it} are vectors of exogenous covariates, and ε_{1it} and ε_{2it} are random components capturing unobserved variables. Z_{it} and X_{it} are observed whenever $a_{it} = 1$, and Z_{i1} and X_{i1} are observed for all individuals in the sample. Similar to Zabel (1998), we include wave dummies in Z_{it} and X_{it} to account for duration dependence. A monotonic change in the coefficients on the wave dummies indicates the presence of such dependence. In Equation (1), negative dependence signifies that the probability of attrition from the survey is increasing over time, *ceteris paribus*. In other words, the likelihood of an individual being observed in the sample decreases over time. On the other hand, positive duration implies that survey participants likely remain as in-scope respondents for the duration of the panel's reference period.

3.3 Wages equation

Wages are given by the following equation:

$$y_{it} = W_{it}\beta + \varepsilon_{3it}, \quad i = 1,\dots, n, \; t = 1,\dots, T_i \tag{3}$$

where y_{it} is log hourly wage,[5] W_{it} is a vector of exogenous covariates, and ε_{3it} is a random component. The structural model is given by Equations (1), (2) and (3). This model is sequential since dummy variable e_{it} is observed only if $a_{it} = 1$ (the individual is in-scope and responds), and y_{it} is observed only if $a_{it} = 1$ and $e_{it} = 1$ (the individual is in-scope, responds and is employed[6]) (see Maddala, 1983, pp. 278-283, for further examples on multiple criteria for selectivity).

Ideally, one would like to estimate the model by considering random terms ε_{jit}, $j=1,2,3$, $t = 1,...,T_i$, to be freely correlated for the same individual. However, doing so will involve computing joint probabilities from a $3xT_i$ variate distribution, which is practically problematic. In order to ease the estimation of the model, we will adopt the random effect model approach (see below). We also estimate the model consistently in two stages following the approach suggested by Ham (1982). The latter approach is an extension of the two-stage estimator for the one selection rule proposed by Heckman (1979), and is computationally more attractive than the maximum likelihood method and produces consistent parameter estimates. The first stage involves a joint estimation of the selection equations (1) and

[5] In the empirical estimation, we consider the composite hourly wage for all paid-worker jobs held by the respondent during year t.

[6] Notice that we estimate Equation (3) using hourly wages, which are given only for paid workers. However, employed workers include non-paid workers. So, the latter are ignored in Equation (3).

(2) using panel data.[7] Then, correction terms using obtained parameter estimates are calculated and inserted in the wages equation (3) to account for selection bias.

3.4 Stage 1: Selection Equations

In order to simplify the computation of joint probabilities, we adopt the random effects model, which specifies:

$$\varepsilon_{1it} = u_{1i} + v_{1it} \qquad \varepsilon_{2it} = u_{2i} + v_{2it} \qquad \varepsilon_{3it} = u_{3i} + v_{3it} \tag{4}$$

where u_{1i} , u_{2i} and u_{3i} are individual specific effects assumed to be freely correlated, but independent of Z_{it} , X_{it} and W_{it} , and of v_{jit} for $j=1,2,3$ and $t = 1,...,T_i$. We also assume that error terms v_{jit} are independently distributed over individuals and time. In addition, v_{jit} are mutually independent. Hence, the correlations between Equations (1), (2) and (3) are given by the correlations between individual specific effects u_{1i}, u_{2i} and u_{3i} . Let $u_i^* = (u_{1i}, u_{2i}, u_{3i})'$ conditional on u_i^* , ε_{1it} , ε_{2it} and ε_{3it} are independent. The vector u_i^* is assumed to follow a trivariate normal distribution:

$$u_i^* \sim N(0, \Sigma) \tag{5}$$

where $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix}$.

Attrition and non-response are random and there is no selectivity bias in employment equation estimates (Equation 2) if unobserved individual determinants of employment are uncorrelated with unobserved determinants of attrition/non-response, (i.e. if $\sigma_{12} = 0$). Likewise, there is no selectivity from attrition when estimating the wage equation if u_{1i} and u_{3i} are uncorrelated (i.e. if $\sigma_{13} = 0$). As described above, the first stage of our procedure involves the joint estimation of Equations (1) and (2). For this purpose, the individual contribution to the likelihood function conditional on

$$u_i = (u_{1i}, u_{2i})' \text{ is: } L_i(u_i) = \prod_{t=1}^{T_i} L_{it}(u_i), \tag{6}$$

where

$$L_{it}(u_i) = \left\{Pr(a_{it} = 0 \mid u_i)\right\}^{1-a_{it}} \times \left[\left\{Pr(a_{it} = 1,\ e_{it} = 0 \mid u_i)\right\}^{(1-e_{it})} \times \left\{Pr(a_{it} = 1,\ e_{it} = 1 \mid u_i)\right\}^{e_{it}}\right]^{a_{it}} \tag{6.1}$$

for $t \geq 2$, and

$$L_{it}(u_i) = L_{i1} = \left\{Pr(e_{it} = 0 \mid u_i)\right\}^{(1-e_{it})} \times \left\{Pr(e_{it} = 1 \mid u_i)\right\}^{e_{it}} \tag{6.2}$$

for $t=1$.

[7] Ham (1982) uses only cross-sectional data.

Since all individuals were in scope and responded to the labour interview at time period t =1, the contribution of an individual to the likelihood function depends only on the employment status at this period. Given that ε_{1it} and ε_{2it} are independent conditional on u_i, Equation (6.1) simplifies to:

$$L_{it}\left(u_i\right)=\left\{Pr\left(a_{it}=0\,|\,u_i\right)\right\}^{1-a_{it}} \times \left[\Pr\left(a_{it}=1\,|\,u_i\right)\left(\Pr\left(e_{it}=0\,|\,u_i\right)\right)^{1-e_{it}}\left(\Pr\left(e_{it}=1\,|\,u_i\right)\right)^{e_{it}}\right]^{a_{it}} \tag{6.3}$$

For identification issues, we shall assume that v_{1it} and v_{2it} are $N\left(0,1\right)$ distributed. Thereafter, the unconditional contribution of an individual to the likelihood function is:

$$L_i=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}L_i\left(u_i\right)g\left(u_i\right)du_{1i}du_{2i} \tag{7}$$

where $g(.)$ is the joint density function of u_{1i} and u_{2i}. Finally, full maximum likelihood estimates of the parameters in (1) and (2) with the homoskedasticity assumption are obtained by maximizing the log likelihood function:

$$log\left(L\right)=\sum_{i=1}^{n}log\left(L_i\right) \tag{8}$$

Since the function in (8) involves two-dimensional integration, direct optimization is generally not feasible. We will use maximum simulated likelihood instead. Notice that the function in (7) is an expectation ($L_i=E_{u_i}\left[L_i\left(u_i\right)\right]$), which can be approximated by a simulated mean:

$$L_{is}=\frac{1}{R}\sum_{r=1}^{R}L_i\left(u_{ir}\right) \tag{9}$$

where u_{ir}, r =1,…,R, are R draws from the bivariate distribution of u_i. u_{1i} and u_{2i} can be specified as linear combinations of two independent $N\left(0,1\right)$, η_{1i} and

$$\eta_{2i}:\ u_{1i}=s_1\eta_{1i}+s_2\eta_{2i} \text{ and } u_{2i}=s_3\eta_{2i} \tag{10}$$

s_1, s_2 and s_3 are three unknown coefficients to be estimated. Notice that u_{1i} and u_{2i} are independent if s_2 = 0. Finally, parameters in (1) and (2) including s_1, s_2 and s_3 are obtained by maximizing the simulated log likelihood:[8]

$$log\left(L_s\right)=\sum_{i=1}^{n}log\left(L_{is}\right) \tag{11}$$

A sample from u_i is constructed as follows. First, we draw two independent samples of size R each from a $N\left(0,1\right)$. Then, a sample from u_i is obtained using formulas in (10). Gourieroux and Monfort (1996) show that if

[8] See Gourieroux and Monfort (1996) and Train (2002) for discussion and statistical background. See also Green (2002) for some applications of the maximum simulated likelihood.

$\sqrt{n} / R \to 0$ and R and $n \to \infty$, then the maximum simulated likelihood estimator and the true maximum likelihood estimator are asymptotically equivalent. In the empirical application, we use R = 50.[9]

3.5 Stage 2: Selection-Adjusted Wages Equation

In the second stage, we estimate the selection-corrected wages equation. The expectation of y_{it} conditional on responding and being employed is (and ignoring correlation across observations):[10]

$$E\left(y_{it} \mid a_{it}^{*} \geq 0,\ e_{it}^{*} \geq 0 \right) = W_{it}\beta - \sigma_{13}\lambda_{1t} - \sigma_{23}\lambda_{2t} \tag{12}$$

Formulas of λ_{1t} and λ_{2t} for t ≥ 2 are given in Ham (1982). For time period t = 1, there is only one source of selection which is employment status; λ_{2t} is simply the inverse Mills ratio. Parameter estimates from the first stage are used to form consistent estimates $\hat{\lambda}_{1t}$ and $\hat{\lambda}_{2t}$ of λ_{1t} and λ_{2t}. Then we estimate β, σ_{13} and σ_{23} by running the pooled OLS regression using the selected sample (as suggested by Wooldrige, 2001, Chapter 17, for a model with one selection criterion):

$$y_{it} = Z_{it}\beta - \sigma_{13}\hat{\lambda}_{1it} - \sigma_{23}\hat{\lambda}_{2it} + \varepsilon_{3it}^{*} \tag{13}$$

where $\varepsilon_{3it}^{*} = \varepsilon_{3it} + \sigma_{13}\left(\hat{\lambda}_{1it} - \lambda_{1it} \right) + \sigma_{23}\left(\hat{\lambda}_{2it} - \lambda_{2it} \right)$

Finally, consistent estimates of the standard errors of the OLS slopes are obtained using formulas from Ham (1982).

4. EMPIRICAL RESULTS

4.1 Non-Attrition and Employment Equations

At the outset, we notice that the estimated coefficient s_2 on η_{2i} in Equation (10) is statistically significant at the 5 percent level, which means that the unobserved individual determinants of non-attrition and employment, u_{1i} and u_{2i}, are correlated. The correlation between these terms is estimated at 0.45, but the estimated correlation between the whole random terms, ε_{1it} and ε_{2it}, is only 0.03, but is statistically significant at the 5 percent level. Hence, estimating non-attrition and employment equations separately will introduce very limited biases.[11] Results suggest also that the individual specific effect is more important on employment than on non-attrition behaviour.

We also conclude that variables that increase work attachment and/or reduce mobility (for instance, education, being married, health status) also increase the likelihood of remaining in scope and responding. In the same vein, being an

[9] We initially estimated the model using R = 30. There is little change in the results when increasing the number of draws from 30 to 50. Nevertheless, given the complexity of the likelihood function and the large size of the sample, the estimation of the model is computationally demanding even with a small number of draws.

[10] We have not found a study in the literature that corrects for two selection sources using panel data. Wooldrige (2001, Chapter 17), presents a case with one selection criterion.

[11] We estimated non-attrition and employment equations separately and we obtained estimates very close to those when the equations are estimated jointly.

immigrant, and especially being an immigrant member of a visible minority group, reduces significantly both the probability of being employed and the probability of remaining a respondent in the sample. Moreover, a person who moved during a year (a signal of geographical mobility) is more likely to be a non-respondent or out of scope in the subsequent year. Results also indicate that females are less likely to attrite, but are less likely to be employed compared to males. On another hand, being a student does not affect survey participation, though it reduces employment. Interestingly, however, age, ownership of the dwelling and family size, have no significant effects on attrition. An important point that is made obvious by our results is that increased family income lowers the likelihood of attrition. This result is in agreement with MaCurdy, Mroz, and Gritz (1998) who find that individuals who exit from the National Longitudinal Survey of Youth (NLSY) come disproportionately from the low income population.

Coefficient estimates on wave dummies support neither positive duration dependence nor negative duration dependence, since these estimates do not change monotonically. The likelihood of attrition is the highest in 2000 and the lowest in 1998, a fact that agrees with descriptive statistics.

4.2 Wage Equations

In interpreting results, we focus on analyzing the biases that arise from ignoring non-attrition and employment selections on the estimation of the wage equation rather than analyzing the effects of covariates on the wage level. It is useful to note that even when the correlation between the non-attrition and employment equations is set to zero, there is almost no effect on the adjusted wage equation estimates, a fact that confirms the mild dependence between the two statuses. The most novel result is that the coefficient estimates on the correction terms are negative and highly significant, indicating the non-randomness of non-attrition and employment behaviours. From Equation (13), we can interpret the negative signs of these estimates as indications that wages are positively correlated with non-attrition and employment. The extent of selection bias from employment is larger than the extent of selection bias from attrition. The earnings gap between the selected (available) sample and a sample drawn randomly with identical observed characteristics is estimated at 9.65% due to non-attrition selection versus 13.51% due to employment selection.[12]

If we ignore the two selection sources, the effects of several covariates on wages are overestimated. For instance, selection adjustment yields to a male-female gap of 17.4% versus 23.1% without such adjustment. Similarly, the gap between a non-immigrant and an immigrant who is not a member of a visible minority group, declines from 4.6% to 2.2%, and from 20.6% to 12.1% between a non-immigrant and an immigrant who is a member of a visible minority group, after the selection adjustment. In addition, returns to education are overestimated, since they decrease significantly after the selection correction.

5. CONCLUDING REMARKS

The structural estimates provide evidence for the non-randomness of the attrition behaviour. The correlation coefficient between random components in non-attrition and employment equations is positive and significant, though small. The wage equation estimates indicate significant selection biases from both non-attrition and employment. Most of the coefficients on covariates are overestimated when selection is ignored. We also conclude that our model estimates when unadjusted over-estimate wages, since wages in the available longitudinal sample likely are higher than wages that would be observed if all respondents initially selected remained in the sample until the end of the panel. Similarly, we find that increased family income lowers the likelihood of attrition, which could result in a further over-estimation for family income.

[12] The gap is calculated by multiplying minus the selection coefficient times the mean value of the correction term.

REFERENCES

Gourieroux, C. and Monfort, A. (1996), "Simulation Based Econometric Methods," Oxford University Press, New York, 1996.

Green, W. H. (2002), "Convenient Estimators for the Panel Probit Model: Further Results," *Department of Economics, Stern School of Business, New York University,*Revised, October, 2002

Ham, J.C. (1982), "Estimation of a Labour Supply Model with Censoring Due to Unemployment and Underemployment," *The Review of Economic Studies* Vol. 49, No. 3 (Jul., 1982), pp. 335-354.

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

Levesque, I. and Franklin, S. (2000), "Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics". Catalogue no. 75F0002MIE2000004. Ottawa: Statistics Canada.

MaCurdy, T., Mroz, T. and Gritz, R.M. (1998), "An Evaluation of the National Longitudinal Survey of Youth," *Journal of Human Resources*, Vol. 33(2), pp. 345-436.

Maddala, G.S. (1983), "Limited Dependent and Qualitative variables in Econometrics," Cambridge University Press, 1983.

Michaud, S. and Webber, M. (1994), "Measuring Non-Response in a Longitudinal Survey: The Experience of the Survey of Labour and Income Dynamics". Catalogue no. 75F0002MIE1994016. Ottawa: Statistics Canada.

Wooldridge, J. M. (2001), "Econometric Analysis of Cross Section and Panel Data," M. Cambridge, Mass: The MIT Press, 2001.

Zabel, J. E. (1998), "An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior" *The Journal of Human Resources*, Vol. 33, No. 2 (Spring, 1998), pp. 479-506.

TABLES AND REFERENCES

The complete version of the manuscript is available upon request by contacting the authors.