Attrition in the Longitudinal Immigration Survey: New Zealand

Paper presented at the New Zealand Association of Economists Conference 2009
at Wellington, New Zealand
1-3 July 2009

John Bryant and Frances Krsinich

Statistics New Zealand
P O Box 2922
Wellington, New Zealand
info@stats.govt.nz

www.stats.govt.nz

Abstract

The Longitudinal Immigration Survey: New Zealand (LisNZ) is a uniquely rich source of data on migrants' labour market and settlement outcomes in New Zealand. But, like all longitudinal surveys, it is subject to attrition. Approximately 14 percent of respondents who were interviewed in wave 1 could not be re-interviewed in wave 2. We investigate whether this attrition will lead to selection bias in typical cross-sectional models using LisNZ data. We apply two closely-related tests developed in labour economics: (i) we examine whether attrition in wave 2 is related to outcomes in wave 1, after controlling for standard explanatory variables; and (ii) we examine whether the relationship between outcome and explanatory variables differs between attritors and non-attritors. Both tests suggest the existence of selection bias. These biases are nevertheless small compared with the size of the coefficients, especially when wages are used as the outcome variable. The small size of the biases means that, at current attrition rates, the LisNZ sample essentially remains representative of its target population.

Introduction

Longitudinal data are essential to addressing key issues in migration policy, such as the relationship between immigrants' labour market performance and time in the host country (Borjas, 1994), or the relationship between initial characteristics and eventual settlement outcomes. However, every longitudinal survey loses some of its original respondents because of factors such as migration, death, and respondent fatigue. Attrition is particularly likely in a mobile population such as recent immigrants. An obvious consequence of attrition is that it reduces sample size, and hence statistical precision. A bigger threat to the usefulness of the data, however, is that respondents who remain in the sample may differ systematically from respondents who attrite. For instance, respondents who remain in a longitudinal migration survey may represent an unusually content or settled subset of all immigrants. Selection bias of this sort makes it difficult to generalise from the data to the population of interest.

The Longitudinal Immigration Survey: New Zealand (LisNZ) provides data on a wide range of personal characteristics and settlement outcomes for a sample of migrants who were approved for New Zealand residence between 1 November 2004 and 31 October 2005. Like all such surveys, LisNZ is subject to attrition. Of the 7,137 respondents who were successfully interviewed at wave 1, 6156 were interviewed at wave 2, giving a wave 1 to wave 2 attrition rate of 14 percent. (LisNZ will have a third wave, but at the time of writing, the data are still being collected and processed.)

In this paper, we estimate the extent to which attrition between waves 1 and 2 of LisNZ has led to selection bias. Our approach is based on that of labour economists such as Becketti et al (1988) and Fitzgerald et al (1998). We investigate biases in simple cross-section models, using four outcome variables: wages, employment status, satisfaction with life in New Zealand, and ownership of a dwelling. We find that attritors and non-attritors are systematically different, in ways that induce selection bias. However, the biases are small enough, and attrition rates low enough, that they have only a minor effect on the representativeness of the data.

Data and methods

Data

The Longitudinal Immigration Survey: New Zealand (LisNZ) follows migrants over the first three years after they are granted, and take up, residence in New Zealand. It covers a wide range of settlement and labour market outcomes. The target population is all migrants (excluding refugees) who were at least 16 years old and were approved for residence in New Zealand from 1 November 2004 to 31 October 2005. Seventy-three percent of respondents were already living in New Zealand at the time residence was granted. Those who were living overseas were eligible for inclusion in the survey if they arrived in New Zealand within 12 months of residence approval. Respondents are interviewed at six months (wave 1), 18 months (wave 2), and 36 months (wave 3).

At wave 1, a sample of 12,202 migrants was randomly selected for inclusion in the survey. Of these, further enquiry showed that 217 were not eligible to take part in the survey, 145 did not arrive in New Zealand in time, and 984 had no initial contact address in New Zealand. Of the remaining 10,856 migrants, 7,137, or 66 percent, were interviewed. The main reason for failing to interview respondents was non-contact. Many migrants could not be found at the addresses they had supplied at the time they were approved for residence. Statistics New Zealand is carrying out further research into the reasons for non-contact, and its implications for the survey, in a companion research project to the one described here. The present paper focuses entirely on attrition between waves 1 and 2. In other words, it focuses on attrition conditional on inclusion in wave 1. As noted earlier, 6,156 of the 7,137 respondents from wave 1 were re-interviewed at wave 2, implying attrition of 14 percent.

LisNZ is a joint project between the New Zealand Department of Labour and Statistics New Zealand. More information on LisNZ is available at the Department of Labour and Statistics NZ websites.

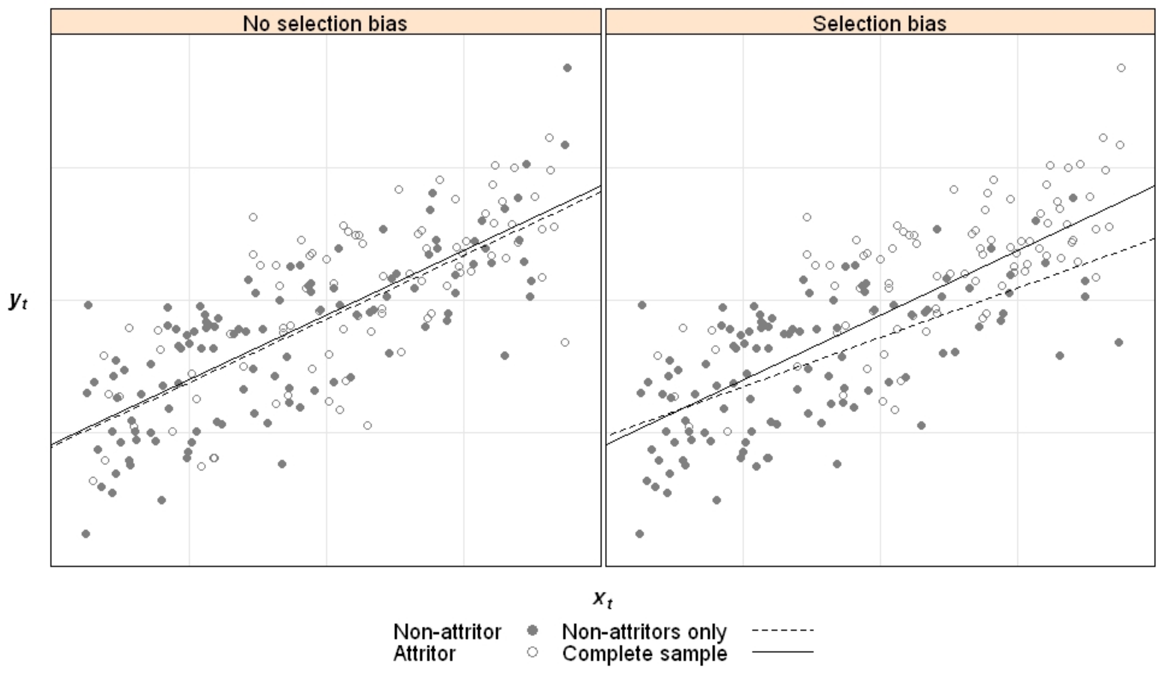
Methods

Setting

We are interested in the relationship between an outcome variable y and a vector of explanatory variables x. Time subscript $t = 1, 2, \cdots$ indexes survey wave. Quantities y_t and x_t typically contain values observed at wave t, but they may also contain values, or functions of values, from previous waves. For instance, y_t could measure changes in wages between waves $t-1$ and t. The survey is subject to attrition, represented by variable A_t. If $A_{it} = 1$ then respondent i attrited in wave t; if $A_{it} = 0$ then the respondent remained in the sample. By definition, there is no attrition at wave 1, so $A_{i1} = 0$ for all i.

Error! Reference source not found. uses two artificial datasets to illustrate the conditions under which attrition leads to selection bias. Each dot in the figure represents a pair of values x_{it}, y_{it}. To keep the graph to two dimensions, vector x_t has been restricted to a single variable. Hollow dots represent respondents who have attrited from the survey at wave t, and solid dots represent respondents who remain. The solid diagonal line is the regression line that would have been obtained if data for the complete sample had been available; the dashed line is the result of regressing on values for non-attritors only.

Figure 1 Two hypothetical datasets illustrating the relationship between attrition and selection bias



In the left panel, attrition does not lead to selection bias. Observations in this panel are “missing at random” and selection is “ignorable” (Little and Rubin, 2002). The slight difference between the regression line for attritors and for the whole sample is due entirely to sampling error. Attrition varies with x_t, but this does not distort the observed relationship between y_t and x_t. In the right panel, attrition does lead to selection bias. Observations are “missing not at random” and selection is “non-ignorable.” The regression line for non-attritors rises less steeply than the line for the complete sample.

The key difference between the left and right panels is that, in the right panel, attrition depends on y_t, even after conditioning on x_t. The importance of dependence on y_t can be seen by applying Bayes’s theorem to the probability density function of y_t conditional on x_t and A_t:

$$f(y_t|x_t, A_t = 0) = \frac{P(A_t = 0|y_t, x_t)f(y_t|x_t)}{P(A_t = 0|x_t)}. \tag{1}$$

Equation 1 implies that

$$f(y_t|x_t, A_t = 0) = f(y_t|x_t), \tag{2}$$

(ie, there is no selection bias) if and only if

$$P(A_t = 0|y_t, x_t) = P(A_t = 0|x_t), \tag{3}$$

(attrition is independent of y_t, after conditioning on x_t). The data in **Error! Reference source not found.** satisfy equation Equation 3 if the ratio of hollow dots to solid dots is independent of position along the vertical axis. The data in the left panel follow this pattern; the data in the right panel do not.

Our objective is to test whether attrition has lead to selection bias: that is, to see whether coefficient estimates obtained from fitting a model to $y_t|A_t = 0$ and $x_t|A_t = 0$ differ from the estimates that would have been obtained if it had been possible to fit the same model to the complete y_t and x_t.

Test 1: The relationship between attrition and the outcome variable

Our first test was developed by Fitzgerald, Gottschalk, and Moffit (1998), under the title of “selection on observables”, and applied by them and, for instance, Alderman et al (2001). Roughly speaking, the test consists of seeing whether the data violate Equation 3. The idea cannot be implemented exactly as stated, since that would require data for attritors in wave t. Additional assumptions must be made, and Equation 3 replaced with an approximation.

The key additional assumptions are that (i) any relationship between A_t and y_t that remains after conditioning on x_t is due to respondent characteristics not captured by x_t, and (ii) these unmeasured characteristics persist between survey waves.[1] One such characteristic might be trust in authority, which is not typically measured, is likely to affect participation in surveys, is likely to affect outcomes such as employment, and is likely to persist over time.

Equation 3 is then approximated by

$$P(A_\infty = 0|y_1, x_1) = P(A_\infty = 0|x_1), \tag{4}$$

where A_∞ measures whether a respondent ever attrites from the survey. Equation 4 is identical to Equation 3, except for the time subscripts. However, the persistent-characteristics assumption implies that changes to the reference period should have only a small effect on results. Testing whether Equation 4 is satisfied should therefore be a good way of testing whether Equation 3 is satisfied.

[1] One way to formalize these ideas is to define a respondent-level fixed effect δ_i such that $y_{it} = f(x_{it}) + \delta_i + \varepsilon_{it}$, $A_{it} = g(x_{it}) + \sigma\delta_i + \gamma_{it}$, and $\mathrm{cov}(\varepsilon_{it}, \gamma_{it}) = 0$.

Equation 4 can be generalized to include data from the first k waves, with A_∞ measuring whether a respondent attrites in waves $k+1$ and higher. The advantage of doing so is that longitudinal models using more than one wave of data can then be accommodated. The disadvantage is that the sample will be missing respondents who attrited in waves 2 to k, reintroducing the possibility of selection bias. This question of whether to include more than one wave is not relevant to us, however, since, at the time of writing, only data from the first two waves of LisNZ are available.

Fitzgerald et al (1998) emphasise that, if x_t consists entirely of variables such as age and sex whose values can be inferred from x_1, and if it is assumed that

$$f(A_\infty = 0|y_t, y_1, x_1) = f(A_\infty = 0|y_1, x_1), \quad (5)$$

then Equation 4 is not merely an approximation of Equation 3, but is equivalent. Fitzgerald et al refer to the situation described in Equation 5, and generalisations of it, as “selection on observables”. They use Equation 5 and the idea of “selection on observables” to clarify the relationship between their approach and that of Heckman (1979) selection models, which, in their terminology, deal with “selection on unobservables”.

If the focus is purely on motivating Equation 4, then Equation 5 and the idea of selection on observables can be safely omitted. Equation 5 cannot be tested, and is not intuitively appealing or obvious. It therefore needs its own justification. The most natural way to do so is to invoke some form of the persistent-characteristics assumption. But if the persistent-characteristics assumption is to be used, then it is simpler to go directly from this assumption to Equation 4, rather than indirectly, via Equation 5.

Equation 4 is tested in practice by regressing A_∞ on y_1 and x_1, and examining the size and statistical significance of the estimated coefficient on y_1. A coefficient that is substantively and statistically different from zero is evidence for selection bias. The test is specific to the y_t, the x_t, and the statistical model being used.

An attractive feature of the approach of Fitzgerald et al (1998) is that it leads naturally to a set of weights for correcting any selection biases that are found. Rearranging Equation 1 gives

$$f(y_t|x_t) = \frac{P(A_t = 0|x_t)}{P(A_t = 0|x_t, y_t)} f(y_t|x_t, A_t = 0) = w(x_t, y_t) f(y_t|x_t, A_t = 0). \quad (6)$$

Equation 6 says that multiplying the distribution of y_t conditional on x_t and $A_t = 0$ by function $w(x_t, y_t)$ gives the distribution of y_t conditional only on x_t. What this means in practice is that weighting each observation x_{it}, y_{it} by quantity $w(x_{it}, y_{it})$ should reduce the amount of selection bias in estimates of the relationship between y_t and x_t. The closer the assumptions underlying Equation 6 are to being met, the greater the reduction in bias.

Test 2: Comparison of distributions

Our second test of whether attrition is leading to selection bias is roughly equivalent to comparing the distribution of the solid dots with the combined distribution of the solid and hollow dots in **Error! Reference source not found.**. The approach is developed in Becketti et al (1988) and applied in, for instance, Fitzgerald et al (1998), Alderman et al (2001), and Anglewicz et al (2009).

The basic idea is to test whether

$$f(y_t|x_t, A_t = 0) = f(y_t|x_t), \quad (7)$$

Once again, this idea cannot be implemented exactly as stated, because values of x_t and y_t for attritors in wave t are, by definition, unknown. The strategy is to approximate equation 7 using

$$f(y_1|x_1, A_\infty = 0) = f(y_1|x_1), \quad (8)$$

and to justify the approximation using some form of persistent-characteristics assumption.

The test is implemented by regressing y_1 on x_1 interacted with A_∞. The results from this regression need to be manipulated, however, before the estimates match the structure of Equation 8. Let $\hat{\beta}_X$ be the main effects from the regression, $\hat{\beta}_{XA}$ the interaction terms, and P_A the proportion of respondents for whom $A_\infty = 1$. The coefficient estimates are then $\hat{\beta}_X$ for non-attritors and $\hat{\beta}_X + \hat{\beta}_{XA}$ for attritors. Estimates for the complete sample are a weighted sum of the two: $(1 - P_A)\hat{\beta}_X + P_A\big(\hat{\beta}_X + \hat{\beta}_{XA}\big)$. Selection bias is measured by the difference between estimates for non-attritors only and for the complete sample, or

$$\hat{\beta}_X - \Big((1 - P_A)\hat{\beta}_X + P_A\big(\hat{\beta}_X + \hat{\beta}_{XA}\big)\Big) = -P_A\hat{\beta}_{XA}. \tag{9}$$

The right hand side of Equation 9 expresses selection bias as the difference between the estimated coefficients for attritors and non-attritors, all multiplied by the attrition rate. The reason the expression has a negative sign is that if the relationship between x_t and y_t is stronger for attritors than for non-attritors, then leaving attritors out biases coefficient estimates downwards.

The test can be expanded to include data from multiple waves, in exactly the same way as the first test. The advantages and disadvantages are the same: the ability to test longitudinal models versus the possibility of introducing selection biases.

The second test complements the first. The first test yields a single estimate (the coefficient on y_1) that can be used as a summary measure of bias from all variables. The second test yields many estimates (the individual components of $-P_A\hat{\beta}_{XA}$), expressed in the units of the model being tested, that can be used to assess the bias associated with each variable.

Description of models

Table 1 summarizes the outcome and explanatory variables used to implement the tests. The outcome variables—employment, wages, satisfaction, and dwelling ownership—are all key components of settlement. The explanatory variables are a standard set that would typically be included in statistical models using LisNZ data.

All of the models, apart from the model of wages in Test 2, are estimated using logistic regression. Following Fitgerald et al (1998), we use the original sample weights provided with the data when estimating the models.

Table 1

Variables used in models

Variable	Description
Outcome variables	
Employed	Dichotomous variable, taking a value of 1 if the respondent was employed or self-employed at the time of the interview, and 0 otherwise
Wages	(Log of) hourly wages for employed respondents
Very Satisfied	Dichotomous variable, taking a value of 1 if the respondent said that he or she was “very satisfied” with life in New Zealand, and 0 otherwise
Owns dwelling	Dichotomous variable, taking a value of 1 if the respondent owns the dwelling where he or she lived at the time of the interview
Explanatory variables	
Application group	Broad immigration application category to which the respondent belongs
Onshore	Dichotomous variable taking a value of 1 if the respondent was living in

	New Zealand at the time his or her residence application was approved
Age	Age in years
Sex	Dichotomous variable taking a value of 1 if the respondent is female and 0 if male
Years of education	Years of education at the time residence was approved
Origin	Respondent's region of origin
English is best language	Whether English is the respondent's best language

Results

Test 1: The relationship between attrition and the outcome variable

Table 2

Attrition as the response variable

Response variable: Attrition = 1								
	employed		logged hourly wages		very satisfied		owns dwelling	
Employed	-0.371	***						
Hourly wages (logged)			0.394	***				
Very satisfied					-0.445	***		
Owns dwelling							-0.460	***
Application group - Business	0.061		0.518	**	0.094		0.120	
Application group - Family	0.048		-0.051		0.065		0.102	**
Application group - Pacific and Other	0.123		0.097		0.138		0.101	
Application group - Skilled (ref)								
Sex - Female	-0.351	***	-0.379	***	-0.299	***	-0.259	***
Onshore	-0.231	***	-0.081		-0.198	***	-0.224	***
Origin - North America & Europe	0.040		0.020		0.026		-0.049	
Origin - North Asia	0.023		0.072		-0.128	*	-0.032	
Origin - Other	-0.293	***	-0.284	**	-0.366	***	-0.439	***
Origin - Pacific	0.080		0.161		0.015		-0.034	
Origin - South Africa	-0.339	***	-0.449	***	-0.366	***	-0.457	***
Origin - South Asia	0.008		0.205	*	-0.108		-0.137	*
Origin - South East Asia	-0.241	**	-0.298	**	-0.363	***	-0.352	***
Origin - UK and Ireland (ref)								
English is best language	-0.059		-0.266	***	-0.091	*	-0.060	
Age	0.009		-0.025		-0.013		0.008	
Age squared	0.000	**	0.000		0.000		0.000	*
Years of education	0.008		-0.008		0.000		0.002	
Intercept	-1.410	***	-1.710	***	-0.988	***	-1.527	***
n	7125		4378		7126		7032	
R square	0.048		0.087		0.058		0.051	

* p < 0.05, ** p < 0.01, ***p<0.001.

Logistic regression was used to estimate all the models.

Table 2 shows the evidence for selection bias for each of employment, wages, satisfaction and dwelling ownership. Each of these variables is significantly associated with attrition.

Those who are employed are less likely to attrite, as are those who are 'very satisfied' with their settlement experience to date, and those who own their own dwelling. Of the employed respondents, the higher their hourly wage the more likely they are to attrite.

The low R squares show that, although the associations between attrition and each of employment, wages, satisfaction and dwelling ownership are statistically significant, the included variables are only explaining a small proportion of the variation in attrition.

Test 2: Comparison of distributions

Table 3

Models including attrition intercept and interactions

	response variable							
	employed		logged hourly amount		very satisfied		owns dwelling	
Application group - Business	-0.392	***	-0.268	***	0.019		1.225	***
Application group - Family	-0.220	***	-0.114	***	-0.052		0.394	***
Application group - Pacific and Other	0.067		-0.164	***	0.239	***	-0.526	***
Application group - Skilled (ref)								
Sex - Female	-1.127	***	-0.163	***	-0.231	***	-0.016	
Onshore	-0.917	***	-0.038	*	-0.279	***	-0.409	***
Origin - North America & Europe	-0.079		-0.049		-0.158	***	-0.678	***
Origin - North Asia	-0.569	***	-0.339	***	-2.285	***	-1.263	***
Origin - Other	0.059		-0.177	***	-0.793	***	-1.641	***
Origin - Pacific	0.104		-0.266	***	-0.681	***	-1.633	***
Origin - South Africa	0.162	**	-0.009		-0.187	***	-1.524	***
Origin - South Asia	-0.251	***	-0.332	***	-1.350	***	-2.205	***
Origin - South East Asia	0.038		-0.230	***	-1.229	***	-1.319	***
Origin - UK and Ireland (ref)								
English is best language	0.443	***	0.076	***	-0.093	**	0.369	***
Age	0.327	***	0.050	***	-0.033	***	0.253	***
Age squared	-0.004	***	-0.001	***	0.000	***	-0.003	***
Years of education	0.093	***	0.043	***	-0.015	***	0.017	***
Intercept	-5.053	***	1.459	***	1.367	***	-6.106	***
Attritor*Application group - Business	-0.077		0.146		-0.628	***	-0.278	
Attritor*Application group - Family	-0.450	***	-0.161	**	-0.046		0.534	***
Attritor*Application group - Pacific and Other	-0.926	***	-0.120		-0.586	***	0.754	*
Attritor* Application group - Skilled (ref)								
Attritor*Sex - Female	-0.418	***	-0.006		0.241	***	0.480	***
Attritor*Onshore	-0.105		0.036		-0.027		-0.268	*
Attritor*Origin - North America & Europe	0.050		-0.081		-0.267	*	0.565	***
Attritor*Origin - North Asia	0.109		-0.040		0.593	***	0.465	**
Attritor*Origin - Other	0.175		0.030		0.543	**	0.349	
Attritor*Origin - Pacific	0.614	***	0.036		0.656	***	-0.239	
Attritor*Origin - South Africa	-0.012		-0.026		-0.558	***	1.130	***
Attritor*Origin - South Asia	0.759	***	-0.029		0.707	***	0.193	

Attritor*Origin - South East Asia	0.093		-0.010		0.182		0.021	
Attritor*Origin - UK and Ireland (ref)								
Attritor*English is best language	-0.455	***	0.010		0.075		-0.433	***
Attritor*Age	0.028		-0.033	*	0.010		0.231	***
Attritor*Age squared	-0.001	*	0.000	*	0.000		-0.003	***
Attritor*Years of education	0.030	*	-0.008		-0.014		0.068	***
Attritor	-0.319		0.746	*	-0.590		-6.541	***
n	7125		4378		7126		7032	
R-square	0.760		0.339		0.482		0.689	

* p < 0.05, ** p < 0.01, ***p<0.001.

Logistic regression was used to estimate all the models except the wages model, which is estimated using Ordinary Least Squares.

Table 3 shows the main effects and interactions of the regressions of $y_1$on x_1 interacted with attrition. That is, the $\hat{\beta}_X$ and $\hat{\beta}_{XA}$ referred to in equation (9).

Attrition is associated with a decreased likelihood of being employed for those in the 'Family' and 'Pacific and Other' application groups, relative to the 'Skilled' application group. Females who are attritors are less likely to be employed than females who are non-attritors. Those respondents from the Pacific and South Asia who attrited are more likely than non-attritors from the same regions to be employed, relative to those from the UK and Ireland. Attritors whose best language is English are less like to be employed than non-attritors whose best language is English.

For the model explaining wages, the intercept term is significant at the 0.05 level, which means that attritors have, on average, higher wages. Most of the attrition interactions, however, are not statistically significant, other than for the interaction with the Family application group, and age. Being an attritor is associated with a decrease in the association with hourly earnings of those in the Family application group relative to the Skilled application group. Attrition is associated with a decrease in the effect of age on hourly wages.

Attritors from the 'Pacific and Other' and Business application groups are less likely to be 'very satisfied' than non-attritors from those groups, relative to the Skilled application group. Female attritors are more likely to be 'very satisfied' than female non-attritors. Attritors from North America, Europe or South Africa are less likely to be very satisfied than non-attritors from the same regions, relative to those from the UK and Ireland. On the other hand, attritors from North and South Asia, the Pacific and 'Other' regions are more likely to be very satisfied than their non-attriting counterparts, relative to those from UK and Ireland.

The intercept term in the model for dwelling ownership is both statistically and practically significant – those who attrite are much less likely to own their own dwelling. Attritors from the Family and the 'Pacific and Other' application groups are more likely to own their own dwellings compared to non-attritors from those application groups, relative to the Skilled application group. The association of being female with dwelling ownership is stronger for attritors than non-attritors. Attritors who are on-shore at the time of residence approval are less likely to own their own dwelling than non-attritors who are on-shore at the time of residence approval. Attritors from North America and Europe, North Asia and South Africa are all more likely to own their own dwelling than non-attritors from the same countries, relative to those from the UK and Ireland. Attritors whose best language is English are less likely to be dwelling owners than non-attritors whose best language is English. Attritors are more likely to own their own dwellings than non-attritors the older they are and the more years of education they have.

Discussion

Attrition between waves 1 and 2 of LisNZ leads to selection biases in simple cross-section models, but these biases are too small to constitute an important threat to the usefulness of the data. The evidence for the existence of selection biases is strong. There are highly statistically significant associations between attrition and important outcome variables, even after conditioning on a standard set of explanatory variables. Moreover, the associations between attrition and outcomes are generally in the expected direction: for instance, respondents who own their own dwelling or who are more satisfied with life in New Zealand are less likely to attrite. Further testing confirms that the relationship between the explanatory and outcomes variables differs between attritors and non-attritors, at least for some variables. However, the differences in coefficients are modest compared to the coefficients themselves. Combined with the fact that attritors only represent 14 percent of all wave 1 respondents, this means that the LisNZ sample effectively remains representative of the target population.

References

Alderman H, Behrman J R, Kohler H-P, Maluccio J A, Cotts-Watkins S (2001). “Attrition in longitudinal household survey data: Some tests for three developing-country samples” *Demographic Research* 5(4): 79-124.

Anglewicz P, Adams J, Obare F, Kohler H-P, Watkins S (2009). “The Malawi diffusion and ideational change project 2004–06: Data collection, data quality, and analysis of attrition” *Demographic Research* 20(21): 503-540.

Becketti S, Gould W, Lillard L, Welch F (1988). “The Panel Survey of Income Dynamics after fourteen years: An evaluation” *Journal of Labor Economics* 6(4): 472-92.

Borjas G J (1994). “The economics of immigration” *Journal of Economic Literature* 32(4): 1667:1717.

Fitzgerald J, Gottschalk P, Moffitt R (1998). “An analysis of sample attrition in panel data: The Michigan Panel Survey of Income Dynamics” *Journal of Human Resources* 33(2): 251-99.

Heckman J J (1979) “Sample selection bias as a specification error” *Econometrica* 47(1): 153-161.

Little R J A, Rubin B (2002). *Statistical Analysis with Missing Data. Second Edition* New York: Wiley