JOINT RESEARCH CENTER FOR PANEL STUDIES
DISCUSSION PAPER SERIES

DP2009-003
August, 2009

Attrition Bias in Longitudinal Survey: A Validation Study with Interviewer's Record of Respondent Mobility

Michio Naoi*

Abstract

This paper aims to examine respondent's mobility-related non-response in the longitudinal survey. We use the interviewer's record of respondent mobility, which can be observed even if the respondent does not participate in the respective wave, as a source of validation data. Using the Keio Household Panel Survey (KHPS) 2004–2007 as a primary dataset, household mobility equations are estimated for selected subsample of non-attritors by two competing methods—an inverse probability weighted (IPW) estimator and a sample selection (SS) estimator. These two estimators are compared with a probit estimates using complete sample including both attritor and non-attritor. It is found that SS generally outperforms IPW in terms of coefficient estimates, suggesting that the mobility-related non-responses in the KHPS are non-ignorable. However, the results of Hausman test cannot find any significant bias for either IPW or SS estimator.

Key words: Attrition Bias, Validation Study, Respondent Mobility, Non-ignorable non-response.

JEL classification: C33, C81, R23.

*Michio Naoi
Faculty of Economics, Keio University

Joint Research Center for Panel Studies
Keio University

Attrition Bias in Longitudinal Survey: A Validation Study with Interviewer's Record of Respondent Mobility*

Michio Naoi†
Keio University

August 21, 2009

Abstract

This paper aims to examine respondent's mobility-related non-response in the longitudinal survey. We use the interviewer's record of respondent mobility, which can be observed even if the respondent does not participate in the respective wave, as a source of validation data. Using the Keio Household Panel Survey (KHPS) 2004–2007 as a primary dataset, household mobility equations are estimated for selected subsample of non-attritors by two competing methods — an inverse probability weighted (IPW) estimator and a sample selection (SS) estimator. These two estimators are compared with a probit estimates using complete sample including both attritor and non-attritor. It is found that SS generally outperforms IPW in terms of coefficient estimates, suggesting that the mobility-related non-responses in the KHPS are non-ignorable. However, the results of Hausman test cannot find any significant biasfor either IPW or SS estimator.

Key words: Attrition Bias, Validation Study, Respondent Mobility, Non-ignorable non-response.

JEL classification: C33, C81, R23.

*The author is grateful to the 21st Century Center of Excellence Program at Keio University for generously providing us with the data (Keio Household Panel Survey). The author would like to thank Colin McKenzie, Koyo Miyoshi, and Miki Seko for their helpful comments and suggestions.

†Faculty of Economics, Keio University, Mita Toho Bldg. 3rd Floor, 3-1-7 Mita, Minato-ku, Tokyo 108-0073. *Email:* naoi@2001.jukuin.keio.ac.jp.

1 Introduction

The primary aim of this paper is to examine respondent's mobility-related non-response in the Keio Household Panel Survey (KHPS) 2004–2007. There are always drop-outs from the survey at each wave and some of them are directly or indirectly related to the respondent's mobility. It is often hard to keep track of the new address of the respondent who moved. Respondent's move pertaining to his/her marriage often leads to survey refusal by his/her new spouse. All these factors cause a close linkage between respondent mobility and survey non-response, and hence potential bias in studying household mobility with longitudinal data.

In the survey sampling literature, a crucial issue in examining non-response in the longitudinal survey is whether or not the underlying missing data process is *ignorable* (Rubin, 1976; Little and Rubin, 2002). Using s as an indicator of survey response ($s = 1$ if respond, 0 otherwise) and y and x as the outcome of interest and other observed characteristics, ignorable non-response or missing-at-random (MAR) can be defined by $\Pr(s = 1|y, x) = \Pr(s = 1|x)$.[1] This implies that, conditional on observed characteristics, survey response behavior is independent of the (possibly unobservable) behavior of interest. In our context, ignorability or MAR requires that the probability of non-response does not vary systematically across movers and non-movers.

Fitzgerald et al. (1998) further extend the notion of ignorability of missing data process by using the concepts of *selection-on-observables* and *selection-on-unobservables*. Introducing an auxiliary variable z, which is distinct from x and always observable, selection-on-observables occurs when $\Pr(s = 1|y, x, z) = \Pr(s = 1|x, z)$. They show that the inverse probability weighting (IPW) — weighting the observed data by the inverse of the probability of response — provides consistent estimates under the selection-on-observable assumption. The selection-on-observable approach requires z to be endogenous to y but excluded from the regression equation. They suggest that the lagged dependent variable can be used as an obvious candidate for z.

On the other hand, selection-on-unobservables, also termed as not-missing-at-random or non-ignorable non-response, occurs when the above conditional probability assumption does not hold. In such a case, an appropriate method to cope with non-response bias follows sample selection model (Heckman, 1979; Hausman and Wise, 1979). In applying a sample selection (SS) model, the identification of the *behavioral* coefficients requires an exclusion restriction, i.e., there should be an instrument z that affects non-response while being independent from the behavior of interest. In practice, however, finding a suitable instrument for unobservable selection is by no means easy in the case of non-response.[2] Parameter estimates from the observed data alone are typically biased to an extent that depends on the strength of the relationship between the unobserved outcome of interest and the probability of non-response.

The non-ignorability assumption, however, is generally untestable because the behavior of interest cannot be observed when the respondent dropped out of the survey. In this paper,

[1] When the underlying missing data process is *completely* at random, $\Pr(s = 1|y, x) = \Pr(s = 1)$ will be satisfied.

[2] Naoi (2007) uses the interview process characteristics as a source of instruments. We also use the same set of IVs in the following analysis.

we use the interviewer's record of respondent mobility to test the non-ignorability assumption. Since the interviewer's record can be observed even if the respondent himself is dropped out from the survey, we can directly test the conditional probability assumption above. In the following analysis, household mobility equations are estimated for both the entire sample including attritors and the selected subsample of non-attritors. The performance of two competing estimators — an inverse probability weighted (IPW) estimator and a sample selection (SS) estimator — is evaluated by comparing them with a probit estimates using complete sample. It is found that SS generally outperforms IPW in terms of coefficient estimates, suggesting that the mobility-related non-responses in the KHPS are non-ignorable. However, the results of Hausman test cannot find any significant bias for either IPW or SS estimator.

The paper is organized as follows. Section 2 briefly introduces the KHPS. Section3 summarizes attrition pattern in the KHPS and presents descriptive analysis of attrition. Section 4 explains the estimation method and variables. Section 5 summarizes our main empirical results. Section 6 presents the conclusion.

2 Data

The KHPS, sponsored by the Ministry of Education, Culture, Sports, Science and Technology, is the first comprehensive panel survey of households in Japan, conducted annually by Keio University since 2004. In the following analysis, we use the first four waves of the KHPS, which were conducted in 2004, 2005, 2006 and 2007, respectively. In 2004, 13,430 individuals, male and female, aged 20–69 years, were selected by stratified two-stage random sampling as a potential respondent. Out of 13,430 individuals initially approached, 4,005 primary respondents finally participated in the first wave of the survey (response rate = 29.8%). Although the overall response rate of the first wave was not so high (29.8%), the age and sex distribution of the initial 4,005 respondents is quite similar to that of the Japanese population.[3]

The questionnaire of the KHPS contains both individual and household related questions. The former covers a wide array of questions with respect to the respondent's demographic characteristics, education and employment activities, while the latter has basic questions pertaining to household income, asset holdings, and housing conditions. If the primary respondent was married at the time of survey, the questionnaire also contains virtually identical questions to be answered by his/her spouse. The standard procedure for the KHPS was to send a pre-survey letter to the respondent and then provide a post-interview payment of 3,000 yen (approximately $25) per household. In addition to the main dataset described above, the interviewer's records are provided about the respondent's moves and the interview process such as the contact history.[4] These supplemental surveys are conducted in 2005, 2006 and 2007 (waves 2 to 4).

[3] Details of data description can be found in Kimura (2005) and Higuchi et al. (2008).

[4] We have also used the complete list of assignments of each interviewer to the targeted respondents in the following analysis. This information was made available by the Central Research Services Inc., to which the author is grateful.

By 2006, the KHPS witnessed a sample loss of approximately 28% due to cumulative attrition from its initial 2004 sample. Compared with other longitudinal surveys, the attrition in the KHPS is somewhat heavy; this can possibly be attributed to the fairly long and comprehensive questionnaire used. For example, Japanese Panel Survey of Consumers (JPSC, cohort A) has 10.1% of sample loss in the initial three waves. The cumulative attrition rates for the first three waves are approximately 15% in the Panel Study of Income Dynamics (PSID) and 5.7% in the National Longitudinal Survey of Youth (NLSY). For the European Community Household Panel (ECHP), these figures range from 12.1% (Germany) to 36.5% (Denmark). In the next section, we will review the general attrition pattern of the KHPS, focusing on the relationship with respondent mobility.[5]

3 Descriptive Analysis of Attrition in the KHPS

Table 1 presents the attrition pattern of the KHPS and its relationship with selected socioeconomic and demographic characteristics of the respondents. The overall drop-outs decrease the initial sample size of 4,005 to 3,314, 2,884, and 2,634, in 2005, 2006, and 2007 respectively.[6] The corresponding drop-out rates for each wave becomes 17.3%, 13.0%, and 8.7%, respectively.[7] As in the most longitudinal surveys, drop-out rate is highest in the second wave (2005), and then becomes lower in the subsequent waves, implying that there may be some duration dependence in the attrition process.

Table 1 also shows that the drop-out rate varies systematically with the respondent's socioeconomic and demographic characteristics at each wave. Except for household mobility, respondent/household characteristics are all measured in wave $t-1$. As expected, respondent's age has clear U-shaped effect on the drop-out probability. In 2005, drop-out rate is lowest of 14.0% for those in their 40s, and highest at the margin, 21.5% for those aged under 30 years and 18.0% for those aged over 60 years. Compared with overall drop-out rates in each wave, female respondents have slightly lower drop-out probability. Respondents who are married or have some college degrees are less likely to be dropped out from the survey. It is also found that those with poor health condition tend to have higher probability of drop-outs. With regard to the relationship between respondent mobility and non-response, Table 1 suggests that the respondent's mobility is strongly and positively related to non-response — in the second wave, 36.3% of movers attrited from the survey as compared to a mere 17.3% of overall drop-out rate. The same pattern can be found in the subsequent waves.[8] Basically, non-response poses serious problems when it

[5] For an extensive review of attrition problem in the KHPS, see Miyauchi et al. (2006), McKenzie et al. (2007) and Naoi (2007).

[6] As in other longitudinal surveys, the KHPS has some individuals who come back into the survey from nonresponse. There are 3 rejoining respondents in the fourth wave (2007). Since the number of rejoining respondents are so small that it is almost negligible, we simply omit them from the following analysis.

[7] The drop-out rate is the percentage of the number of drop-outs in wave t to the number of respondents in wave $t-1$.

[8] Information on residential mobility used here is based on the interviewer's record, which can be obtained even if the respondent is dropped out from the survey. "Recent mover" in Table 1 represents those who have moved

is not independent from the behavior of interest, suggesting that sample attrition is especially important when we try to understand the household residential mobility.

(Table 1 around here)

Table 2 presents summary statistics of first-wave respondent/household characteristics by subsequent survey responses status. To create this table, we use the same set of individuals as in the empirical analysis conducted in section 5. If the primary respondent is married and is female, individual characteristics (age, years of education, labor force participation, and health condition) are for her male spouse. The left panel shows summary statistics for two groups of respondents — those respond and drop out in the second wave. The right panel shows summary statistic calculated for those participate all four waves of the KHPS, and those drop out in either waves 2–4. Generally, we can observe the same pattern of survey response as shown in Table 1. Non-response decreases with individuals' age at the start of the panel. Non-response is greater among those with less formal education and with poor health condition. Among others, marital status and household type are found to be a strong predictor of survey response; those respond in the second wave are likely to be married and less likely to be in a single family household in the first wave. These points will be further investigated in Table 4 below.

(Table 2 around here)

For those dropped out from the survey at each wave, Table 3 reports major reason for non-response. Approximately 30% of attritors in each wave reported that they did not participate in the survey because they were too busy. Length of survey questionnaire also matters, and the proportion of individuals choosing this reason is closely related to the actual volume of questionnaire. The proportion was 24.6% and 25.1% in 2005 and 2006 where the questionnaire had 56 pages, and 21.1% in 2004 where the volume of questionnaire reduced to 47 pages. Feeling distrustful about the KHPS/survey in general constitutes over 20% of reason for non-response in 2005. The proportion, however, becomes smaller in the subsequent waves, suggesting that, on the one hand, individuals less willing to participate in the survey dropped out in the earlier waves, and, on the other, ongoing participation may form some kind of "familiarity" or "trust" to the survey itself. Proportions of individuals choosing other reasons are relatively stable across waves.

(Table 3 around here)

To extend the descriptive analysis of non-response presented in Tables 1 and 2, Table 4 presents probit models for survey response. The dependent variable for these models equal 1 if the individual responds at the wave in question and 0 otherwise. The probability of response is modelled as a function of the wave $t-1$ values of regressors, which can be observed regardless of the survey response status in wave t, and wave dummies. Table 4 shows the marginal effects of

between waves $t-1$ and t.

the regressors on the probability of response at each wave. The marginal effects are evaluated at the sample means. The marginal effects for dummy variables represent a discrete change from zero to one.

Model [1] is our baseline result. The results reveal statistically significant associations between non-response and respondent's age. As expected, respondents who are married or have higher educational status are likely to remain in the survey. Respondents from the single family household have significantly lower retention probability. Wave dummies are included to account for possible duration dependence, and all of them are highly significant with monotonically increasing marginal effects. This suggests that there is strong evidence of positive duration dependence in the KHPS.[9]

In model [2], we additionally introduce two explanatory variables of household mobility in order to test the ignorability, or selection-on-observability, assumption presented in section 1, i.e. $\Pr(s = 1|y, x, z) = \Pr(s = 1|x, z)$. These mobility indicators are obtained from the interviewer's record, hence they can be observable even if the respondent attrites from the respective waves of the panel. Move_t represents the behavior of interest — whether or not the respondent moves between waves $t-1$ and t — which is denoted by y in the above conditional probability. Following previous studies, we use the lagged variable of y (Move_{t-1}) as the auxiliary variable z (Fitzgerald et al., 1998; Jones et al., 2006). Selection-on-observability requires that, once controlling for z, y should be independent of the response, thereby it does not have any direct effect on the survey response. It is found that the wave t mobility (Move_t) has strongly negative effect on the wave t response probability even after controlling for lagged mobility indicator (Move_{t-1}). This implies that, in the KHPS, the ignorability assumption does not hold for the analysis of household mobility. Therefore we expect that an appropriate method to account for non-response bias will be sample selection models rather than inverse probability weighting. In the next section, we will explain our empirical methods (SS and IPW) in detail, and compare the estimated mobility equations by these two competing methods in section 5.

(Table 4 around here)

4 Models and Estimation Methods

Consider a longitudinal survey that includes T waves. Attrition occurs if individual i leaves the survey in period $T_i \leq T$. If information for individual i is not observed for any wave after T_i, attrition is an absorbing state. Attrition is nonabsorbing if an individual can return to the sample after exiting. Attrition is considered to be absorbing for this analysis. Thus, the attrition

[9] Positive duration dependence arises when the survey response becomes less time consuming as the respondents repeatedly participate in the survey, or the respondents feel some form of 'familiarity' or 'trust' to the interviewer/survey. Since any unobserved heterogeneity causes spurious duration dependence, we also estimate the model including the first wave values of the all time-varying variables (Zabel, 1998). This yields qualitatively same results as in Table 4.

process is specified as

$$s_{it} = X_{i,t-1}\alpha_1 + D_{it}\beta_1 + Z_{it}\gamma_1 + \varepsilon_{1it} \quad i = 1, \cdots, N, t = 1, \cdots, T_i \tag{1}$$

where

$$\begin{aligned} \text{individual } i \text{ remains in wave } t \ (s_{it} &= 1) \text{ if } s^*_{it} > 0 \\ \text{leaves in wave } t \ (s_{it} &= 0) \text{ if } s^*_{it} \leq 0. \end{aligned}$$

$X_{i,t-1}$ is a $1 \times M$ vector of regressors, D_{it} is a $1 \times T$ vector of wave dummies, Z_{it} is a $1 \times K$ vector of auxiliary variables/instruments, and ε_{1it} is an independent and identically distributed (i.i.d.) normal random variable with mean zero and variance σ_1^2 that is independent of X_{it-1}.[10]

Our interest of household mobility equation is modeled along with the attrition equation. Consider the following panel data model of household mobility behavior.

$$y_{it} = X_{i,t-1}\alpha_2 + D_{it}\beta_2 + \varepsilon_{2it} \quad i = 1, \cdots, N, t = 1, \cdots, T_i^* \tag{2}$$

where

$$\begin{aligned} \text{individual } i \text{ moves between waves } t-1 \text{ and } t \ (y_{it} &= 1) \text{ if } y^*_{it} > 0 \\ \text{does not move between waves } t-1 \text{ and } t \ (y_{it} &= 0) \text{ if } y^*_{it} \leq 0. \end{aligned}$$

We use the same set of regressors of $X_{i,t-1}$ and D_{it} as in equation (1). ε_{2it} is i.i.d. normal random variable with mean zero and variance σ_2^2. Since the observed move (y_{it}) takes place between waves $t-1$ and t, household residential mobility is modelled as a function of the wave $t-1$ values of regressors, which can be observed regardless of the survey response status in wave t. Note that, in the usual situation, respondent moves (y_{it}) cannot be observed if individual i does not rameins in the wave t, because they are surveyed at wave t panel. $T_i^* = T_i - 1$ in this case. In our current situation, however, interviewer's record of respondent mobility can be observed even at $t = T_i$. Hence $T_i^* = T_i$ in our current situation.

In the following analysis, we will first estimate equation (2) using entire sample by probit model. We will then estimate equation (2) for selected subsample of non-attritors (i.e. excluding observations in $t = T_i$ for those with $T_i < T$) by two estimation methods explained below.

4.1 Inverse Probability Weighted Estimator

To compute the IPW estimator, we first estimate probit model for survey response (equation 2) using lagged mobility indicator as an auxiliary variable (i.e. $Z_{it} = y_{i,t-1}$). Using the estimated

[10] If the household mobility decision follows a dynamic utility maximization problem, the individual fixed effects, which is a function of all information available at $t = 1$, will exist (Heckman and Macurdy, 1980). This implies that ε_{2it} is correlated with $X_{i,t-1}$. Following Zabel (1998), we model the individual fixed effects as function of wave 1 values of time-varying variables.

result, the fitted response probability in wave t, conditional that the individual i is in the wave $t-1$ panel, can be obtained. Denoting this conditional probability as $\hat{\pi}_{it}$, the predicted probability weights are constructed cumulatively using $\hat{p}_{it} = \hat{\pi}_{i2}\hat{\pi}_{i3}\cdots\hat{\pi}_{it}$. Finally, using $1/\hat{p}_{it}$ as a weight for observations in selected subsample of non-attritors, equation (2) is estimated by probit ML.

4.2 Sample Selection Estimator

Using selected subsample of non-attritors, we employ the method proposed by Van de Ven and Van Pragg (1981), which extends Heckman's (1979) selection model to the case of a probit model with sample selection. For the identification of the parameters, we assume that ε_{1it} and ε_{2it} have unit variances and are joint normally distributed. Setting $\rho = \text{Corr}\,(\varepsilon_{1it}, \varepsilon_{2it})$, the null hypothesis of interest is given by $H_0 : \rho = 0$, which indicates that there are no attrition biases. This can be tested either by the Wald or likelihood ratio tests.

The key variable in equation (1) is the identifying instrument (Z_{it}) which has been excluded from equation (2). To construct this variable, we use information about the interviewer assignment. The supplemental questionnaire of the KHPS provides detailed information to the targeted respondents about the complete list of assignments of each interviewer. Using this information, we construct the following three variables for z_{it}: (1) the number of respondents (including the targeted respondent) which the interviewer is in charge of, and (2) whether or not the assigned interviewer is the same as the one in the previous wave, and (3) the wave 2 retention rate of the targeted respondents for each interviewer. The first two variables are used in Naoi (2007), and the third variable is used to control for the quality of each interviewer.

5 Empirical Results

The estimation results for equation (2) are presented in Table 5. In the first column, probit result with entire sample (including both attritors and non-attritors) is presented as our baseline result. To assess the extent of the attrition bias for each parameter estimate, results of (1) probit model using subsample of non-attritors, (2) IPW model, and (3) sample selection model are presented along with the baseline model.[11]

(Table 5 around here)

Comparing the results of three alternative models with that of baseline model, we find that the regression coefficients generally show similar signs although there are a number of sizable differences in magnitude and significance. For example, while health condition and length of stay appear to be significant predictor of residential moves in the baseline model, estimated coefficients for these two variables are not estimated to be significant in the three alternative models. In terms of the magnitude of the estimated coefficients, relative differences of the estimated coefficients

[11] Results of probit models for survey response can be available upon request.

from probit with entire sample are presented in Table 6. Denoting parameter estimate of the baseline model as θ_0 and the alternative model as $\tilde{\theta}$, relative difference is defined by $|\theta_0 - \tilde{\theta}|/|\theta_0|$. It is found that sample selection model generally outperforms IPW and uncorrected probit model in terms of coefficient estimates, which is consistent with our findings shown in Table 4. Although there are sizable differences for several variables, coefficient estimates of the sample selection model are virtually identical with baseline case for "single family," "labor force participation," and "housing tenure."

(Table 6 around here)

To examine the overall size of non-response bias, Hausman tests of the hypothesis $\theta = \theta_0$ are conducted for three alternative models. The results of test are presented in the bottom of Table 5. Standard implementation of this test requires estimation of $\mathrm{V}(\theta_0 - \tilde{\theta})$, and obtaining this estimate can be difficult without the strong assumtions on the covariance structure. Here, we use the paired bootstrap to obtain consistent estimate of $\mathrm{V}(\theta_0 - \tilde{\theta})$ (Cameron and Trivedi, 2005, p.378). The results indicate that the null hypothesis of identical coefficients can be rejected at the 10% significant level only for uncorrelated probit model. Although our descriptive analysis of non-response suggest that the ignorability assumption does not hold for the analysis of household mobility, and that we expect IPW model cannot adequately account for non-response bias, Hausman test cannot find significant bias in the the estimated coefficients of IPW model even at the 10% significance level.

6 Conclusion

This paper aims to examine respondent's mobility-related non-response in the longitudinal survey. We use the interviewer's record of respondent mobility, which can be observed even if the respondent does not participate in the respective wave, as a source of validation data. Using the Keio Household Panel Survey (KHPS) 2004–2007 as a primary dataset, household mobility equations are estimated for selected subsample of non-attritors by two competing methods — an inverse probability weighted (IPW) estimator and a sample selection (SS) estimator. These two estimators are compared with a probit estimates using complete sample including both attritor and non-attritor. It is found that SS generally outperforms IPW in terms of coefficient estimates, suggesting that the mobility-related non-response in the KHPS is non-ignorable. However, the results of Hausman test cannot find any significant bias for either IPW or SS estimator.

References

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Fitzgerald, J., P. Gottschalk, and R. Moffit (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources 33*(2), 251–299.

Hausman, J. A. and D. A. Wise (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica 47*(2), 455–474.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J. and T. E. Macurdy (1980). A life cycle model of female labour supply. *Review of Economic Studies 47*(1), 47–74.

Higuchi, Y., M. Kimura, and M. Naoi (2008). Keio Household Panel Survey (KHPS): Outline and purposes. In M. Yano (Ed.), *The Japanese Economy — A Market Quality Perspective*, Chapter 2, pp. 21–30. Tokyo: Keio University Press.

Jones, A. M., X. Koolman, and N. Rice (2006). Health-related non-response in the British Household Panel Survey and European Community Household Panel: Using inverse-probability-weighted estimators in non-linear models. *Journal of Royal Statistical Society A 169*(3), 543–569.

Kimura, M. (2005). Design and sample characteristics of the 2004 Keio Household Panel Survey (KHPS). In Y. Higuchi (Ed.), *Dynamism of Household Behavior in Japan [I]*, Chapter 1. Tokyo: Keio University Press.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

McKenzie, C. R., M. Naoi, T. Miyauchi, and K. Kiso (2007). Attrition and individual behavior in the labor market. In Y. Higuchi and M. Seko (Eds.), *Dynamism of Household Behavior in Japan [III]*, Chapter 1, pp. 13–75. Tokyo: Keio University Press. (In Japanese).

Miyauchi, T., C. R. McKenzie, and M. Kimura (2006). An analysis of sample attrition and survey response behavior in panel data. In Y. Higuchi (Ed.), *Dynamism of Household Behavior in Japan [II]*, Chapter 1, pp. 9–52. Tokyo: Keio University Press. (In Japanese).

Naoi, M. (2007). Residential mobility and panel attrition: Using the interview process as identifying instruments. *Keio Economic Studies 44*(1), 37–47.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Van de Ven, W. P. and B. M. Van Pragg (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics 17*(2), 229–252.

Zabel, J. (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. *Journal of Human Resources 33*(2), 479–506.

Table 1: Drop-Out Rates by Selected Socioeconomic and Demographic Characteristics, KHPS

	Wave 2 (2005)	Wave 3 (2006)	Wave 3 (2007)
All Data			
Number of respondents	3,314	2,884	2,634
%drop-outs	17.3%	13.0%	8.7%
Age			
20s	21.5%	19.0%	10.1%
30s	16.8%	10.6%	9.2%
40s	14.0%	11.8%	7.7%
50s	17.2%	12.2%	7.2%
60s+	18.0%	13.9%	10.3%
Gender			
Female	17.0%	13.0%	7.7%
Marital Status			
Married	15.6%	11.6%	8.3%
Education			
Some college or above	16.1%	12.9%	8.4%
Health Condition			
Poor health	26.7%	20.0%	13.3%
Housing Tenure			
Homeowner	17.1%	12.8%	8.2%
Private renter	19.1%	15.6%	10.5%
Public renter	11.0%	10.6%	7.9%
Mobility			
Recent mover	36.3%	32.5%	26.2%

Notes: Initial number of respondents in the first wave (2004) was 4,005. Respondent/household characteristics are all measured in wave t - 1. Mobility indicator is obtained from the interviewer's record, where "recent mover" is those who have moved between waves t - 1 and t.

Table 2: First-wave Respondent/Household Characteristics by Subsequent Survey Responses

	Averages			
	Response in Wave 2		Responses in Waves 2 - 4	
Variables	In	Out	Always in	Out
Age	47.56	47.09	47.81	46.79
Years of education	13.20	13.00	13.21	13.06
Marrital status (1 = married)	75.6%	66.0%	77.1%	67.1%
Household Type				
Single family (# of HH member = 1)	7.5%	12.5%	6.8%	11.7%
Nuclear family (1 < # of HH member < 5)	71.3%	66.8%	71.6%	68.1%
Extended family (# of HH member >5)	21.3%	20.6%	21.6%	20.2%
Labor force participation (1 = worked in the last month)	84.3%	81.5%	84.7%	81.9%
Health condition (1: good - 5: poor)	1.97	1.94	1.95	1.99
Housing tenure (1 = homeowner)	75.6%	74.1%	75.9%	74.1%
Length of stay in the current residence	15.76	16.47	15.61	16.47
Place of residence				
14 major cities	24.1%	25.2%	24.8%	23.3%
Other cities	57.6%	56.6%	57.4%	57.5%
Town/village	18.3%	18.2%	17.8%	19.2%
N	2,701	606	2,235	1,072

Notes: Respondent/household characteristics are all measured at wave 1 (2004). If the primary respondent is married and is female, individual characteristics (age, years of education, labor force participation, and health condition) are for her male spouse.

Table 3: Major Reasons of Non-response to the Survey

Reason	Wave 2 (2005)	Wave 3 (2006)	Wave 3 (2007)
Too busy	28.0%	30.4%	34.0%
Too long questionnaire	24.6%	25.1%	21.1%
Feel meaningless to answer the same questionnaire as in the previous wave	23.7%	24.8%	30.5%
Distrustful to the KHPS	10.9%	4.8%	3.7%
No specific reasons	10.5%	13.1%	11.8%
Others	10.5%	13.5%	12.5%
Too difficult questionnaire	9.8%	2.7%	4.5%
Distrustful to surveys in general	9.6%	6.3%	2.8%
Health concern	8.7%	11.4%	11.8%
Privacy issue	8.4%	6.8%	3.7%
Disagreement by the family members	8.0%	8.2%	8.7%
Personal reasons	4.0%	4.3%	3.7%
No feedback from the survey organizer	0.3%	0.0%	0.0%
Insufficient reward	0.0%	0.0%	0.9%

Table 4: Probit Models for Survey Responses

Dependent variable:	Model [1]			Model [2]		
Survey response (1 = respond in wave t)	dF/dx	(S.E.)		dF/dx	(S.E.)	
Age (× 100)	0.0747	(0.0352)	*	0.0515	(0.0352)	
Years of education	0.0027	(0.0016)	+	0.0028	(0.0016)	+
Marrital status (1 = married)	0.0948	(0.0612)	+	0.0894	(0.0602)	+
Household Type						
Single family (# of HH member = 1)	-0.0401	(0.0206)	*	-0.0326	(0.0201)	+
Nuclear family (1 < # of HH member < 5)	-0.0008	(0.0092)		0.0012	(0.0092)	
Extended family (# of HH member >5)	(omitted category)			(omitted category)		
Labor force participation (1 = worked in the last month)	0.0223	(0.0113)	*	0.0192	(0.0111)	+
Health condition (1: good - 5: poor) (× 100)	-0.0667	(0.3521)		-0.1713	(0.3504)	
Housing tenure (1 = homeowner) (× 100)	0.4475	(0.9614)		-0.9427	(0.9397)	
Length of stay in the current residence (× 100)	-0.0695	(0.0307)	*	-0.0694	(0.0306)	*
Place of residence						
14 major cities	0.0149	(0.0109)		0.0146	(0.0109)	
Other cities (× 100)	0.0090	(0.0099)		0.0089	(0.0099)	
Town/village	(omitted category)			(omitted category)		
Wave dummies						
Wave 2	(omitted category)			(omitted category)		
Wave 3	0.0616	(0.0072)	**	0.0605	(0.0072)	**
Wave 4	0.0977	(0.0070)	**	0.0972	(0.0069)	**
Household mobility						
Move_{t-1} (1 = move between waves t - 2 and t - 1)	---			0.0115	(0.0178)	
Move_{t} (1 = move between waves t - 1 and t)	---			-0.1871	(0.0285)	**
Log likelihood	-3093.321			-3060.485		
N	8,415			8,415		

Notes: Marginal effects are evaluated at the sample means of the regressors. **, *, and + indicate that the estimated marginal effect is significant at the 0.01, 0.05, and 0.10 levels, respectively. Except for household mobility, all regressors are measured at wave t - 1. Mobility indicators are obtained from the interviewer's record. "Move_t" equals 1 if the respondent moves between waves t - 1 and t, and "Move_{t-1}" is the lagged variable for "Move_t." The wave 1 marital status is also controlled in the model but omitted from the results.

Table 5: Estimation Results for Household Mobility

Dependent variable:	Probit with Entire Sample		Probit with Non-attritor		Inverse Probability Weighting		Sample Selection	
Household mobility (1 = move between waves t - 1 and t)	Coef.	(S.E.)	Coef.	(S.E.)	Coef.	(S.E.)	Coef.	(S.E.)
Age	-0.0168	(0.0029) **	-0.0159	(0.0034) **	-0.0165	(0.0036) **	-0.0148	(0.0033) **
Years of education	0.0175	(0.0133)	0.0238	(0.0153)	0.0251	(0.0155)	0.0245	(0.0149)
Marrital status (1 = married)	-0.5519	(0.3107) +	-0.7709	(0.3324) *	-0.7503	(0.3200) *	-0.6986	(0.3172) *
Household Type								
Single family (# of HH member = 1)	0.4026	(0.1367) **	0.4463	(0.1638) **	0.4682	(0.1797) **	0.4026	(0.1588) *
Nuclear family (1 < # of HH member < 5)	0.1831	(0.0889) *	0.2228	(0.1052) *	0.2336	(0.1054) *	0.2181	(0.1029) *
Extended family (# of HH member >5)	(omitted category)		(omitted category)		(omitted category)		(omitted category)	
Labor force participation (1 = worked in the last month)	-0.2283	(0.0963) *	-0.2497	(0.1130) *	-0.2779	(0.1180) *	-0.2308	(0.1094) *
Health condition (1: good - 5: poor)	-0.0720	(0.0301) *	-0.0465	(0.0347)	-0.0458	(0.0349)	-0.0451	(0.0337)
Housing tenure (1 = homeowner)	-0.7923	(0.0676) **	-0.8149	(0.0781) **	-0.8000	(0.0810) **	-0.7963	(0.0765) **
Length of stay in the current residence	-0.0059	(0.0033) +	-0.0033	(0.0037)	-0.0029	(0.0038)	-0.0037	(0.0036)
Place of residence								
14 major cities	-0.0359	(0.1011)	-0.0645	(0.1181)	-0.0807	(0.1213)	-0.0546	(0.1148)
Other cities	0.0041	(0.0916)	-0.0077	(0.1068)	-0.0109	(0.1088)	-0.0039	(0.1038)
Town/village	(omitted category)		(omitted category)		(omitted category)		(omitted category)	
Wave dummies								
Wave 2	(omitted category)		(omitted category)		(omitted category)		(omitted category)	
Wave 3	0.0887	(0.0731)	0.0645	(0.0832)	0.0567	(0.0828)	0.0080	(0.0811)
Wave 4	-0.0666	(0.0796)	-0.0945	(0.0896)	-0.1097	(0.0904)	-0.1149	(0.0878)
Constant	-0.7215	(0.2763) **	-1.0852	(0.3223) **	-1.0649	(0.3291) **	-1.2024	(0.3138) **
f(ρ)	---		---		---		2.2330	(1.1750) *
Hausman test statistics / (P-value)	---		21.7	(0.086) +	17.2	(0.246)	14.7	(0.400)
Log likelihood	-1053.155		-767.808		-773.208		-3712.444	
Number of observation	8,415		7,343		7,343		8,415	
Number of censored observation	---		---		---		1,072	

Notes: **, *, and + indicate that the estimated coefficient is significant at the 0.01, 0.05, and 0.10 levels, respectively. All regressors are measured at wave t - 1. The wave 1 marital status is also controlled in the model but omitted from the results. f(ρ) = 1/2*ln[(1 + ρ)/(1 - ρ)]. Standard error for f(ρ) and the variance-covariance matrix for Hausman test statistic are computed by a non-parametric bootstrap procedure with 200 replications.

Table 6: Coefficient Differences from Probit with Entire Sample

Dependent variable:	Probit with Entire Sample	Probit with Non-attritor	IPW	SS
Household mobility (1 = move between waves *t* - 1 and *t*)	Coef.	% Difference	% Difference	% Difference
Age	-0.0168 **	5.9%	2.2%	12.0%
Years of education	0.0175	36.1%	43.3%	39.6%
Marrital status (1 = married)	-0.5519 +	-39.7%	-35.9%	-26.6%
Household Type				
Single family (# of HH member = 1)	0.4026 **	10.9%	16.3%	0.0%
Nuclear family (1 < # of HH member < 5)	0.1831 *	21.7%	27.6%	19.1%
Extended family (# of HH member >5)	---	---	---	---
Labor force participation (1 = worked in the last month)	-0.2283 *	-9.4%	-21.7%	-1.1%
Health condition (1: good - 5: poor)	-0.0720 *	35.4%	36.4%	37.4%
Housing tenure (1 = homeowner)	-0.7923 **	-2.9%	-1.0%	-0.5%
Length of stay in the current residence	-0.0059 +	43.8%	51.4%	37.8%
Place of residence				
14 major cities	-0.0359	-79.5%	-124.5%	-52.0%
Other cities	0.0041	-288.4%	-368.3%	-195.4%
Town/village	---	---	---	---