CLS

Centre for Longitudinal Studies

CLS Cohort Studies

Working Paper 2010/6

Nonresponse Weight Adjustments Using Multiple Imputation for the UK Millennium Cohort Study

John W. McDonald
Sosthenes C. Ketende

November 2010

Nonresponse Weight Adjustments Using Multiple Imputation for the UK Millennium Cohort Study

John W. McDonald
and
Sosthenes C. Ketende[1]

Centre for Longitudinal Studies
Institute of Education, London

November 2010

[1] John W. McDonald, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (John.McDonald@ioe.ac.uk); Sosthenes C. Ketende, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (S.Ketende@ioe.ac.uk)

First published in November 2010 by the
Centre for Longitudinal Studies
Institute of Education, University of London
20 Bedford Way
London WC1H 0AL
www.cls.ioe.ac.uk

ISBN: 978-1-906929-21-3

The Centre for Longitudinal Studies (CLS) is an ESRC Resource Centre based at the Institution of Education. It provides support and facilities for those using the three internationally-renowned birth cohort studies: the National Child Development Study (1958), the 1970 British Cohort Study and the Millennium Cohort Study (2000). CLS conducts research using the birth cohort study data, with a special interest in family life and parenting, family economics, youth life course transitions and basic skills.

The views expressed in this work are those of the authors and do not necessarily reflect the views of the Economic and Social Research Council. All errors and omissions remain those of the authors.

This document is available in alternative formats.
Please contact the Centre for Longitudinal Studies.
tel: +44 (0)20 7612 6875
email: info@cls.ioe.ac.uk

Contents

Abstract

This paper discusses nonresponse weight adjustments for sweep 3 of the UK Millennium Cohort Study (MCS). Weight adjustments are available for monotone patterns of nonresponse, where the nonresponse weight is the inverse of the estimated probability of response based on a logistic regression model, which uses data from previous sweeps to predict response at the current sweep. For non-monotone patterns, some cases have missing data for previous sweeps and this approach cannot be easily applied. For MCS, 7.5% of the families took part in sweeps 1 and 3, but not sweep 2, i.e., a non-monotonic pattern of nonresponse for 1,444 families.

Our approach to estimate a nonresponse weight for MCS sweep 3 was to use multiple imputation to impute the required missing values at sweep 2 for these 1,444 families for the logistic model for response at sweep 3. This imputation used information from sweeps 1 and 3 and only involved imputing the missing values for time-varying variables shown to be predictive of nonresponse in MCS. This resulted in the multiple imputation of nonresponse weights at sweep 3, which can be averaged to produce a single nonresponse weight or the 10 imputed nonresponse weights can used for separate analyses and the results combined using Rubin's rules. We discuss the advantages and disadvantages of both approaches.

Acknowledgements

The authors would like thank G. Kalton for helpful references and J. N. K. Rao for comments.

Key Words

Imputation; Longitudinal survey; Wave nonresponse; Weighting adjustment.

1 Introduction

1.1 Wave nonresponse in longitudinal surveys

A common problem in longitudinal studies such as panel and birth cohort studies is wave nonresponse, which occurs when responses are obtained for some, but not all, waves of the study. Little and David (1983) discuss three types of wave nonresponse: attrition, re-entry and late entry. Attrition occurs when a unit drops out of the study at one wave and remains out thereafter; re-entry occurs when a unit drops out for one or more waves, but re-enters the study at a later point; and late entry occurs when a unit does not participate in the first wave, but enters the study later. At wave 3 of the UK Millennium Cohort Study (MCS), the longitudinal pattern of response is complex, with attrition, re-entry and late entry patterns of response. At wave 3, 2,210 families attrited at wave 1 and 1,664 attrited at wave 2; 692 families entered the MCS late at wave 2 and of these “new families” 124 attrited at wave 2; and 1,444 families re-entered the MCS by participating at wave 1, not at wave 2, but re-entered the MCS at wave 3.

The pattern of longitudinal participation may be monotone, where no participant returns to the study after missing a wave, or non-monotone, where some subjects return to the study after missing a wave. This paper discusses unit nonresponse weight adjustments for wave (sweep) 3 of the MCS. Weight adjustments are available for monotone patterns of wave nonresponse, where the nonresponse weight is the inverse of the estimated probability of response at the current wave based on a logistic regression model, which uses data from previous waves to predict response at the current wave. For non-monotone patterns, some cases have missing data for previous waves and this approach cannot be easily applied. For the MCS, the longitudinal pattern of response is complex and non-monotonic with families that enter late at wave 2 and re-enter at wave 3 after not participating at wave 2. Our approach to estimate nonresponse weights for MCS wave 3 was to use multiple imputation to impute the required missing values at wave 2 for the logistic model for response at wave 3. This imputation used information from waves 1 and 3 and only involved imputing the missing values for time-varying variables shown to be predictive of nonresponse in MCS. This resulted in the multiple imputation of nonresponse weights, which can be averaged to produce a single nonresponse weight or the 10 imputed nonresponse weights can used for separate analyses and the results combined using Rubin’s rules. In this paper, we discuss the advantages and disadvantages of both approaches.

1.2 Compensating for wave nonresponse in panel surveys

Kalton (1986) and Lepkowski (1989) review methods for compensating for wave nonresponse in panel surveys. Both discuss using weighting or imputation and the reasons why as a rule unit nonresponse is treated by weighting and item nonresponse is treated by imputation. Which strategy is used often depends upon the auxiliary information available for use in making the nonresponse adjustments.

Weighting is often favoured when the auxiliary variables available are only weakly related to the variables with missing values, e.g. when only survey design information is available. Imputation is often favoured when the auxiliary variables available are strongly related to the variables with missing values. In the case of longitudinal surveys, two general imputation approaches are possible: cross-sectional imputations and cross-wave imputations. Cross-sectional imputations are imputations based on data from a single wave. Cross-wave imputations are imputations based on data from multiple waves. For example, for a two-wave panel with repeated variables, one would use the value of a variable on one wave to impute the missing value of the same variable on the other wave; if the missing value was for the second wave, this would be forecasting and if the missing value was for the first wave, this would be backcasting. Imputing for wave nonresponse for variables which are repeated at each wave may be the best solution if the responses for the repeated variable are highly correlated over time. In general, missing values may be forecast using values of the variable from previous waves or backcast using values of the variable from subsequent waves or `interpolated' if the missing wave is surrounded by adjacent non-missing waves. Of course, cross-wave imputation schemes would also typically use other variables as well as the cross-wave repeated variables in the imputation model.

Since both weighting and imputation each have their strengths (advantages) and weaknesses (disadvantages), both Kalton (1986) and Lepkowski (1989) consider that a combination of weighting and imputation methods may be the best solution. This combination strategy was adopted in this paper to estimate wave nonresponse weight adjustments. As our desired logistic regression model for response at wave 3 included predictor variables with missing data, we used cross-wave multiple imputation to deal with the item missing data on some of the predictor variables. This resulted in the multiple imputation of nonresponse weights at wave 3.

1.3 Nonresponse weight adjustments for the UK Millennium Cohort Study

The UK Millennium Cohort Study population is children born between 1 September 2000 and 31 August 2001 (for England and Wales), and between 24 November 2000 and 11 January 2002 (for Scotland and Northern Ireland), alive and living in the UK at age 9 months, and eligible to receive Child Benefit at that age; and then, after 9 months, for as long as they remain in the UK at the time of sampling. Previous Great Britain/United Kingdom birth cohort studies in 1946, 1958 and 1970 sampled all births in one week. In contrast, the MCS sampled births spread across a calendar year and used a complex sample design. The MCS was a stratified one-stage cluster sample of electoral wards. The UK population was stratified by country - England, Wales, Scotland and Northern Ireland. Using a ward-level measure of child poverty and ethnicity, wards in England were stratified into three strata labelled advantaged, disadvantaged and ethnic and wards in Wales, Scotland and Northern Ireland were stratified into two strata labelled advantaged and disadvantaged. A different proportion was sampled from each of the 9 strata, so that the smaller UK countries are over-represented relative to England, and families living in more disadvantaged areas are over-represented and in England, families living in areas with high proportions of ethnic minority families are also over-represented. Once the 398 wards were selected, a list of all 9 month old children living in these wards was required. The lists were generated by using government Child Benefit records. Child Benefit is a universal provision, payable from the child's date of birth, and Child Benefit claims cover virtually all of the UK population, except those who do not have a `right to reside' in the UK. All families, who were eligible to receive child benefit and whose child was living in the sampled ward at age 9 months, were invited to participate in the MCS via a letter from the government department which administered the Child Benefit records. The response rate for MCS wave 1 was 72%. For further details on the MCS sample design and implementation, see Plewis (2007a).

Families moved in and out of the selected sample wards. Families can continue to be paid Child Benefit without notifying the government of a change of address, especially if the benefit was paid directly into a bank account. As a result, the list of children residing in the selected wards was out of date and some eligible children were missed as their existence in the selected ward was not known during the wave 1 fieldwork. The wave 2 sample was supplemented by 692 "new families" eligible at wave 1, but missed because their addresses were not up to date. The cohort child in new families entered the MCS late at age 3.

The productive sample at wave 1, i.e., with interview data from at least the main respondent or partner, formed the issued sample at wave 2, except for cohort members known to have died or emigrated. At wave 2, wave 1 refusals were not issued to the fieldwork agency. The response rate for wave 2 of the MCS was 79%. The correlates of unit nonresponse in the first two waves of MCS are described in Plewis (2007b) and Plewis et al. (2008). Plewis (2007b) found for MCS waves 1 and 2 that differences in response probabilities were small compared to the unequal selection probabilities in the sample design.

Plewis (2007b) describes the methodology used to generate the MCS wave 2 nonresponse weights. A set of potential predictors of nonresponse at wave 2 was studied using a logistic regression model. They included a range of socio-demographic and socio-economic explanatory variables along with a measure of residential mobility derived from the address database, which is part of the administrative side of the MCS survey operation (see Plewis et al. (2008) for more details of this measure). The new families who entered MCS late were excluded from this modelling and their nonresponse weights at wave 2 were defined to be one. All the explanatory variables (other than residential mobility) were measured at wave 1. Because of the issuing strategy, i.e. refusers were not issued, there was no missing data in the wave 1 variables used in the modelling. The predicted probabilities of responding based on this logistic regression model were inverted to generate the nonresponse weights at wave 2. The overall weights at wave 2 are the product of the wave 1 overall weights and the wave 2 nonresponse weights.

For wave 3, all families who were not interviewed at wave 2 were re-issued to the fieldwork agency, except for permanent refusals and families where the cohort member was known to have died or emigrated. Many families not interviewed at wave 2 were productive at wave 3. This yielded a non-monotonic pattern of longitudinal response with 1,444 families that were productive in waves 1 and 3, but not wave 2. As these 1,444 cases have missing data for wave 2 and the estimation of nonresponse weights using the logistic regression approach cannot be easily applied unless all the explanatory variables in the model were measured at wave 1. While some of the explanatory variables Plewis (2007b) found to be predictive of response from wave 1 to wave 2 were time constant such as the ethnic group of the cohort member, others were time varying such as family income, housing tenure (own, rent, other) and type of accommodation (house/bungalow, other). We would expect that the more recent values at wave 2 would be more predictive of response at wave 2, than the values at wave 1, and the values of these time varying explanatory variables are missing at wave 2 for the 1,444 families that were productive in waves 1 and 3, but not wave 2. However, one might think that this is the ideal situation for using imputation for the wave 2 missing values as we have the values of these variables at adjacent waves.

2 Imputation of missing wave 2 data

Multiple imputation was used to replace the missing wave 2 data with "plausible" values. Note that we are not imputing the entire wave, but only the missing values for time-varying variables shown to be predictive of nonresponse in MCS by Plewis (2007b). Multiple imputation is used to take account of the uncertainty in the imputation process. We impute multiple times, in our case 10 times, to create 10 datasets, where the missing values have been filled in. We use each of the 10 datasets to fit our logistic model for response at wave 3 using the same set of explanatory variables. Each dataset yields, typically different, estimates of the nonresponse weights for wave 3 of the MCS. So we end up with 10 estimates of the nonresponse weights. Various issues arise: 1) how should the imputations be carried out, 2) what are the assumptions of the imputation procedure and whether these assumptions justified and 3) how does the analyst deal with multiple estimates of the nonresponse weights.

2.1 Multivariate imputation by chained equations

Multivariate imputation by chained equations (MICE) is the name of the procedure for imputing missing multivariate data by fully conditional specification (van Buuren, 2007). The basic idea is to impute the missing values of variables on a variable-by-variable basis by specifying one imputation model per variable. For each missing variable, a univariate conditional distribution for the missing variable, given other variables, can be specified, i.e., the imputation model is specified separately for each variable, involving the other variables as predictor variables. The method consists of iterating over these conditional distributions by means of Gibbs sampling. At each stage of the algorithm, an imputation is generated for the missing variable and this imputed value is used in the imputation of the next variable. Each iteration cycles through all the variables with missing data. This process is repeated until the process reaches convergence. For details, see van Buuren (2007). MICE has been implemented in a number of software packages including Stata, R and SPSS (Horton and Kleinman 2007, van Buuren and Groothuis-Oudshoorn forthcoming). We used Patrick Royston's implementation of MICE using his Stata ado files (Royston, 2004; Royston, 2005a; Royston, 2005b; Royston, 2007).

In theory, multiple imputation works, but in practice using MICE with Stata 10 was problematic as most of our variables were categorical variables and we had a large dataset. Having interactions in an imputation model meant we had to explicitly create a dummy variable for each category of categorical variables with more than two categories. The dummy variables were imputed when missing and then the original categorical variable was passively imputed from its imputed dummy variables. Issues we had to deal with in practice included predictors, which perfectly predicted binary variables so that we had to fix errors interatively by selecting a reference category other than the first or collapsing categories. Dealing with interactions in imputation models takes care (von Hippel 2009), in our case mainly because of sparse tables. For reasons of space we will not further discuss the details of the imputation models used, other than to say that we used cross-wave imputations and tried to include the sampling design in our imputation models by including the stratum variable as one of our predictors (Reiter et al., 2006). We were not able to account for clustering in our imputation models, but our logistic model for response included the stratum variables as predictors and took account of the clustered nature of the MCS sample. Note that Stata 11 has a new mi (multiple imputation) command, which resolves many of the above issues we had using Stata 10.

Our imputation model makes the assumption that the unobserved values are MAR (missing at random), i.e., given the observed data, the missingness mechanism does not depend on the unobserved data. In the context of a longitudinal study with a monotone pattern of nonresponse, the assumption that the missing longitudinal survey data are MAR is usually invoked when estimating nonresponse weights, so that the probability of response at wave t is assumed to be a function of variables measured at previous waves, but not on unobserved variables measuring changes between wave t-1 and t. Plewis et al. (2008) shows that residential mobility, i.e., any change of postal address, after wave 1 was an important predictor of nonresponse at wave 2 in the MCS, which casts doubt on this common assumption. Our nonresponse model for MCS wave 3 did include residential mobility between wave 2 and wave 3 as a predictor variable.

3 Handling multiply imputed nonresponse weights

3.1 Pooling the results using Rubin’s rules

Multiple imputation fills in each missing value with a set of M plausible values to generate M completed datasets, in our case M = 10 datasets. Each of these 10 datasets were used to estimate 10 nonresponse weights for MCS wave 3, which were used in the standard way to produce cross-sectional or longitudinal weights. These weights can used to produce a weighted estimate of some quantity of interest, say a proportion, a mean or the regression coefficient in a logistic regression model. The results may be combined or pooled, using what have been termed Rubin’s rules (Rubin, 1987), to give estimates and standard errors that take into account the uncertainty due to the imputed missing data values in the nonresponse model used to estimate the nonresponse weights.

3.2 Alternatives

One alternative is to use the mean of the M estimated nonresponse weights for analysis. Use of the mean weight has the following advantages: 1) it is simpler to use the mean weight with a single dataset rather than having use M datasets and then combine the results using Rubin’s rules and 2) it is simpler to deposit one dataset in a data archive than M datasets. Use of the mean weight has the following disadvantages: 1) it does not take into account all the uncertainty in the results due to the incomplete data, 2) it is more complicated to deal with M datasets, both for the user, data provider and data archivist, and 3) users need to understand multiple imputation and be trained to use relevant software which deals with M datasets and pools the results.

Rather than invert the M estimated response probabilities from the logistic model for response at wave 3 and average, an alternative is to average the M estimated response probabilities and then invert. Note that the harmonic mean is always less than or equal to the arithmetic mean, with equality only when all numbers are equal. This implies that for each case, the mean weight is always greater than or equal to the inverse of the mean probabilities. As these two quantities are equal only when all M estimated response probabilities are equal. These two quantities will vary with the variability of the M estimated response probabilities. It is not clear which measure is to be preferred on statistical grounds.

3.3 Is it necessary to use multiply imputed nonresponse weights?

Is it necessary to use multiply imputed nonresponse weights with M datasets and pool the results? What are the differences if we used a single dataset and the mean of the weights or the inverse of the mean of the estimated response probabilities to weight each observation? We investigated these issues by estimating a mean, a proportion and a logistic regression coefficient using all three alternatives. At MCS wave 3, we estimated the mean weight for boys in kilograms, the proportion of boys overweight or obese and the regression coefficient for residential mobility in a logistic model for response at wave 3. Figures 3.3-1 to 3.3-3 present from left to right, a pooled estimate using the 10 datasets and Rubin's rule for pooling the results, estimates based on each of the 10 datasets, denoted IM1, IM2, …, IM10, an estimate using the mean weight and an estimate using the inverse of the mean of the estimated response probabilities. For the limited cases studied, the estimate using Rubin's rules and the mean weight gave very similar estimates, with the lowest estimate in all cases being for the inverse of the mean of the predicted response probabilities.

Figure 3.3-1: Estimated mean weight for boys in kilograms at Millennium Cohort Study wave 3

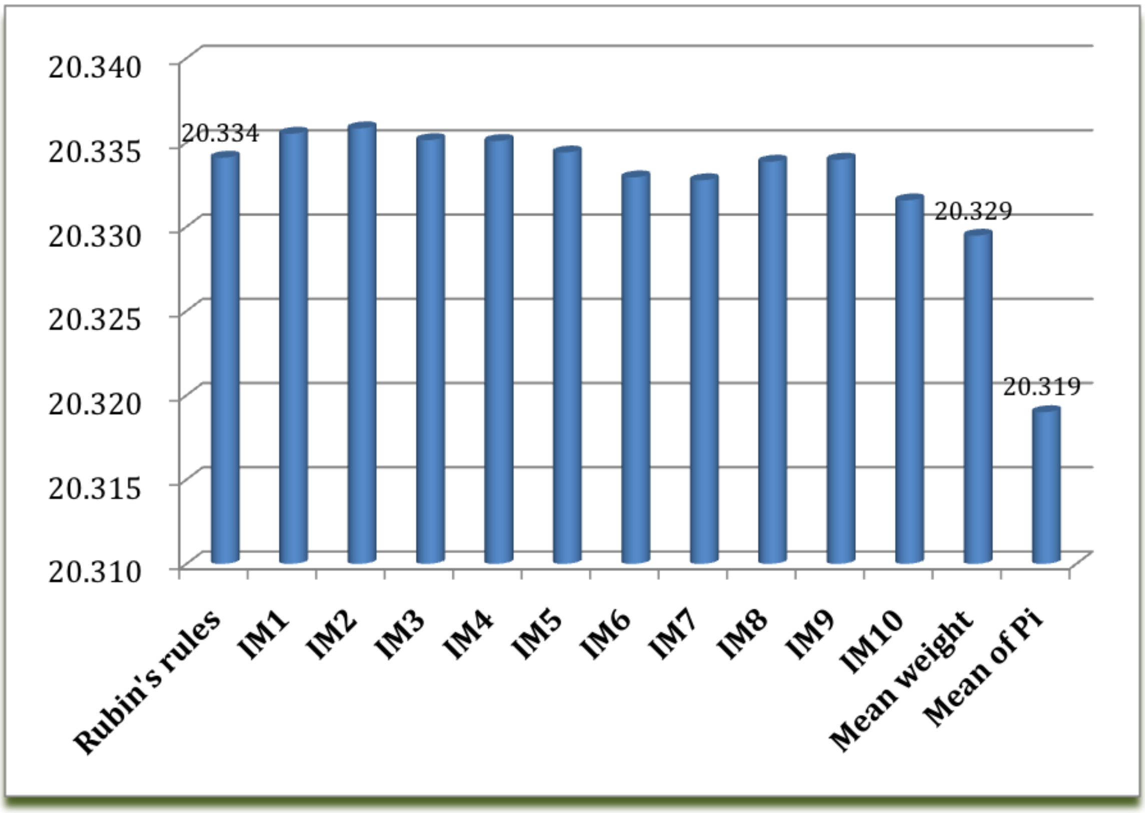


Figure 3.3-2: Estimated proportion overweight or obese at Millennium Cohort Study wave 3

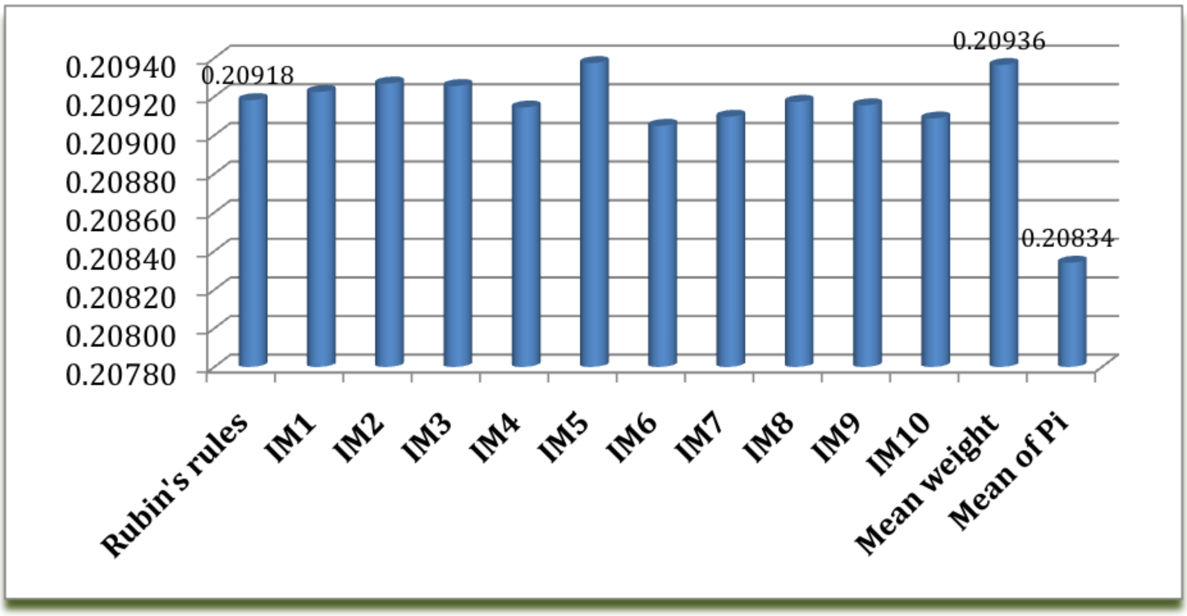
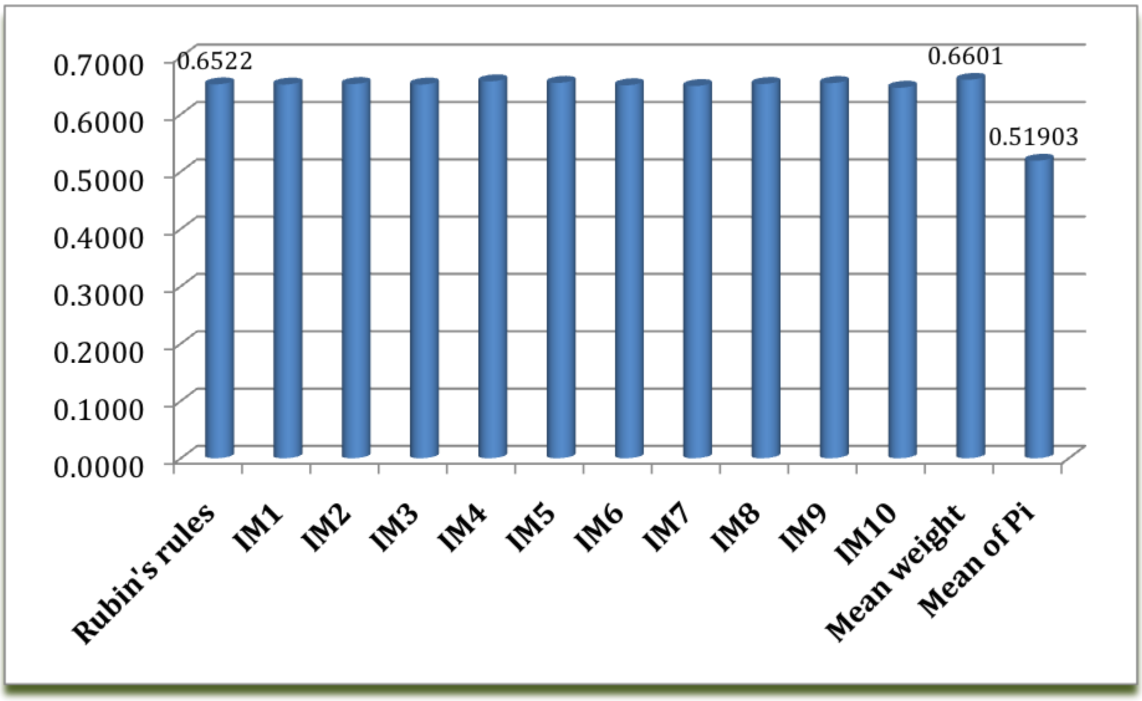


Figure 3.3-3: Estimated regression coefficient for residential mobility in logistic model for response at Millennium Cohort Study wave 3



References

Horton, N. J., and K. P. Kleinman (2007), “Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Regression Models”, *The American Statistician,* 61, pp. 79-90.

Kalton, G. (1986), “Handling Wave Nonresponse in Panel Surveys”, *Journal of Official Statistics*, 2, pp. 303-314.

Lepkowski, J. M. (1989), “Treatment of Wave Nonresponse in Panel Surveys”, in D. Kasprzyk et al. (eds*.) Panel Surveys*, New York: Wiley, pp. 348-374.

Little, R. J. A., and M. H. David (1983), “Weighting Adjustments for Nonresponse in Panel Surveys”, Working paper, U. S. Bureau of the Census, Washington, D. C.

Plewis, I. (ed.) (2007a). “Millennium Cohort Study First Survey: Technical Report on Sampling (4th ed.) ”, London: Centre for Longitudinal Studies.

Plewis, I. (2007b), “Nonresponse in a Birth Cohort Study: The Case of the Millennium Cohort Study”, *International Journal of Social Research Methodology,* 10, pp. 325-334.

Plewis, I., S. C. Ketende, H. Joshi, and G. Hughes (2008), “The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study”, *Journal of Official Statistics* 24, pp. 365-385.

Reiter, J. P., T. E. Raghunathan, and S. Kinney (2006), “The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data”, *Survey Methodology*, 32, pp. 143-150.

Royston, P. (2004), “Multiple Imputation of Missing Values”, *Stata Journal*, 4, pp. 227-241.

Royston, P. (2005a), “Multiple Imputation of Missing Values: update”, *Stata Journal*, 5, pp. 188-201.

Royston, P. (2005b), “Multiple Imputation of Missing Values: Update of ice”, *Stata Journal*, 5, pp. 527-536.

Royston, P. (2007), “Multiple Imputation of Missing Values: Further Update of ice, With an Emphasis on Interval Censoring”, *Stata Journal*, 7, pp. 445-464.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

van Buuren S., H. C. Boshuizen, and D. L. Knook (1999), “Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis”, *Statistics in Medicine*, 18, pp. 681-694.

van Buuren, S. (2007), “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification”, *Statistical Methods in Medical Research*, 16, pp. 219–242.

van Buuren, S., and K. Groothuis-Oudshoorn (forthcoming), “MICE: Multivariate Imputation by Chained Equations in R”, *Journal of Statistical Software*.

von Hippel, P. T. (2009), “How to Impute Interactions, Squares, and Other Transformed Variables”, *Sociological Methodology*, 39, pp. 265–291.

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@ioe.ac.uk
Web http://www.cls.ioe.ac.uk