DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 35, ARTICLE 38, PAGES 1135−1148
PUBLISHED 20 OCTOBER 2016
http://www.demographic-research.org/Volumes/Vol35/38/
DOI: 10.4054/DemRes.2016.35.38

Descriptive Finding

Multiple imputation for demographic hazard models with left-censored predictor variables: Application to employment duration and fertility in the EU-SILC

Michael S. Rendall

Angela Greulich

Contents

Demographic Research: Volume 35, Article 38
Descriptive Finding

Multiple imputation for demographic hazard models with left-censored predictor variables: Application to employment duration and fertility in the EU-SILC

Michael S. Rendall[1]

Angela Greulich[2]

Abstract

OBJECTIVE
A common problem when using panel data is that individuals' histories are incompletely known at the first wave. We demonstrate the use of multiple imputation as a method to handle this partial information, and thereby increase statistical power without compromising the model specification.

METHODS
Using EU-SILC panel data to investigate full-time employment as a predictor of partnered women's risk of first birth in Poland, we first multiply imputed employment status two years earlier to cases for which employment status is observed only in the most recent year. We then derived regression estimates from the full, multiply imputed sample, and compared the coefficient and standard error estimates to those from complete-case estimation with employment status observed both one and two years earlier.

RESULTS
Relative to not being full-time employed, having been full-time employed for two or more years was a positive and statistically significant predictor of childbearing in the multiply imputed sample, but was not significant when using complete-case estimation. The variance about the 'two or more years' coefficient was one third lower in the multiply imputed sample than in the complete-case sample.

[1] Department of Sociology and Maryland Population Research Center, University of Maryland, College Park, USA. E-Mail: mrendall@umd.edu.
[2] Department of Economics, Université Paris 1 Panthéon Sorbonne, France. E-Mail: Angela.Greulich@univ-paris1.fr.

CONTRIBUTION
By using MI for left-censored observations, researchers using panel data may specify a model that includes characteristics of state or event histories without discarding observations for which that information is only partially available. Using conventional methods, either the analysis model must be simplified to ignore potentially important information about the state or event history (risking biased estimation), or cases with partial information must be dropped from the analytical sample (resulting in inefficient estimation).

1. Introduction

A frequently encountered problem when using panel data for demographic applications is that the individual's history is incompletely known at the first wave. It is common practice, but inadvisable (Özcan, Mayer, and Luedicke 2010), to ignore this left censoring. We suggest that multiple imputation (MI), a method typically used to handle non-response (Johnson and Young 2011), can be a general solution to the problem of left censoring in demographic hazard modeling. The Missing at Random (MAR) assumption needed for multiple imputation (Little and Rubin 2002) is sometimes problematic for data that is missing because of non-response (Allison 2001). The MAR assumption is much less likely to be problematic in the case of left censoring, however, as the missingness occurs "by design" (Raghunathan and Grizzle 1995). The 'design' in the case of panel surveys refers to the start date of the panel.

In a previous treatment of this problem in the US Panel Study of Income Dynamics, Moffitt and Rendall (1995) used a maximum likelihood approach to combine left-censored and non-left-censored spells of single motherhood in separate components of the likelihood. The statistical equivalence of maximum likelihood and multiple-imputation (MI) approaches to handling missing data has been noted (Schafer and Graham 2002; White and Carlin 2010). This equivalence assumes 'congenial' imputation and analysis models in which the variables used in the imputation model are also those used in the analysis model, and that the number of imputations $m \rightarrow \infty$ (Meng 1994: 543–544). Moreover, m need not be very large (Schafer and Graham 2002). Our choice of $m = 20$ in the present study reflects the relatively high proportion of person-year cases that are 'incomplete' (almost 50%).

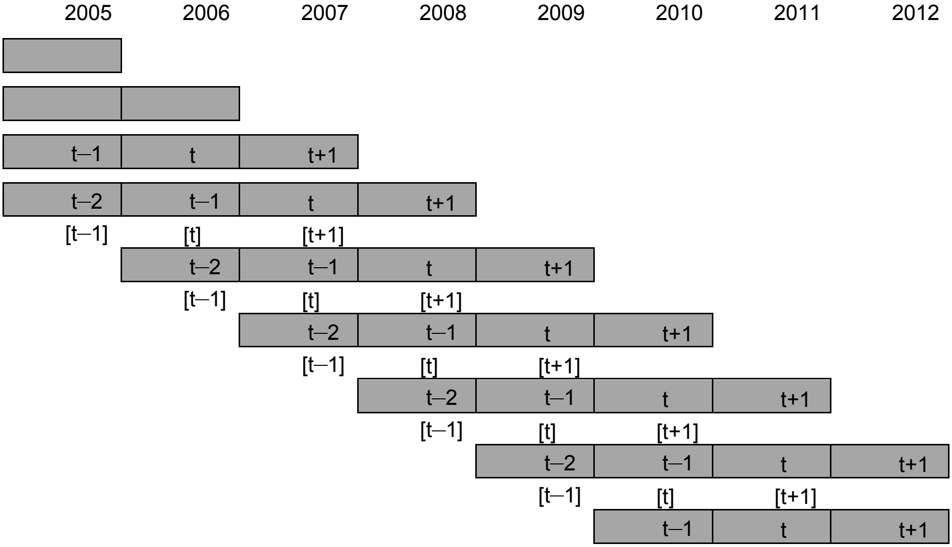
Separating the imputation step from the analysis step, however, has the major advantage of allowing the analysis step to use software designed for rectangular data structures. To our knowledge, MI has not previously been used to address the left-censoring problem. We evaluate gains to using MI this way in a simple example in which women in a panel survey contribute either one or two waves of employment status as a predictor of partnered first birth.

2. Data and method

We use the Poland country sample of the European Union Survey of Income and Living Conditions (EU-SILC), a group of more than 30 comparable country surveys (Eurostat 2011). The standard longitudinal implementation of the EU-SILC consists of a rotational panel in which individuals are observed annually for a period of four years, with four rotation groups present in each year. Poland's SILC follows this design. Selection into the sample occurs annually, beginning in 2005, and each new sample after 2005 is followed for four waves (see Figure 1). We choose Poland due to its relatively large number of observations, reflecting its larger population size than most EU-SILC countries, and because its wave-to-wave retention rate of approximately 90% is among the highest of the EU-SILC countries (Iacovou, Kaminska, and Levy 2012).

We limit our analyses to partnered women aged 18 to 39 who were observed for either three or four consecutive waves between the years 2005 and 2012 and who were childless when entering the survey. The upper age restriction is needed so that we may reasonably approximate a woman's being of parity 0 from a household variable of her having no co-resident children. The restriction to individuals observed for a minimum of three waves is necessary since interviews usually take place during the first half of each year and children born in the second half of each year are then reported at the interview of the year after the birth. Two consecutive years of interviews are therefore needed to identify all births that occur in one calendar year, and at least one more preceding wave is needed to observe the woman's employment status before exposure to conception and birth. For those individuals who are observed for three waves, the latter two waves are designated t and $t+1$ and serve to identify a first birth interval in the calendar year of wave t. Wave $t-1$ is used to observe the woman's employment status and that of her partner. For those individuals who are observed for four waves and for whom no birth occurs in the calendar years of the first two waves, the woman's employment status before birth exposure in year t is observed in both waves $t-2$ and $t-1$.

Figure 1: 'Complete' and 'incomplete' person-year sequences in the Poland country sample of the European Union Survey of Income and Living Conditions, 2005 to 2012



Note: 'Complete' person-year sequences include observation at t–2; 'incomplete' sequences do not.

2.1 Multiple Imputation (MI) for left-censored observations

We specify a model with outcome variable Y_t for a birth in calendar year t as a function of predictor variables observed at times $t-2$ and $t-1$. We allow woman's full-time employment status E to have an effect on Y_t based on its values at both times $t-2$ and $t-1$, E_{t-2} and E_{t-1}. The other predictor variables, denoted by Z and consisting of age (operationalized as 'age – 18') and partner's employment status, have effects on Y_t only from their values at time $t-1$. We delete observations with item non-response, which is anyway very low for our variables of interest in the Poland EU-SILC. That is, no imputation is attempted for missingness due to non-response, which we have noted above is more problematic with respect to the MAR assumption. This leaves us with N_1 'complete' person-year observations $\{Y_t, E_{t-2}, E_{t-1}, Z_{t-1}\}_{i=1}^{N_1}$, omitting person-year subscripts throughout, and N_2 'incomplete' observations $\{Y_t, E_{t-1}, Z_{t-1}\}_{j=1}^{N_2}$. Figure 1 illustrates the five types of complete observations and the seven types of incomplete observations in our data. Whether a woman's birth-exposure year is preceded by one or two years of observed employment status depends on when she was sampled into the panel. Therefore employment status two years before birth exposure is reasonably treated as missing at random (MAR).

We first use the set of complete observations to estimate an imputation equation for E_{t-2}. We use sequential MI (Raghunathan et al. 2001) that allows for the imputation of binary, count, or continuous variables. In our case, the imputed variable is binary, and therefore logistic regression is appropriate:

$$LOGIT[\Pr\{E_{t-2} = 1 | E_{t-1}, Z_{t-1}, Y_t\}] = \gamma_0 + \gamma_1 E_{t-1} + \gamma_2 Z_{t-1} + \gamma_3 Y_t \qquad (1)$$

We then apply random draws from the posterior distribution of parameter estimates $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3$ to the incomplete data $\{E_{t-1}, Z_{t-1}, Y_t\}_{j=1}^{N_2}$ to derive an arbitrarily large number of values m of E_{t-2} (we set $m = 20$) to produce completed data $\{\{Y_t(k), E_{t-2}(k), E_{t-1}(k), Z_{t-1}(k)\}_{j=1}^{N_2}\}_{k=1}^{m}$. Following that, we concatenate the complete data $\{Y_t, E_{t-2}, E_{t-1}, Z_{t-1}\}_{i=1}^{N_1}$ to each instance of completed data and estimate the analysis equation m times. These m estimates are combined using standard multiple-imputation algorithms, or "combining rules" (Little and Rubin 2002), to produce a set of parameters with standard errors that adjust for the uncertainty introduced by imputation of E_{t-2} to the incomplete person-year observations. These combining rules account for the additional uncertainty due to imputation by adding 'between imputation' variance to 'within imputation' variance, thereby avoiding the underestimation of variance of single-imputation analysis (Zhang 2003: 584). This sequence of procedures is performed with standard package software SAS PROC MI and PROC MIANALYZE (SAS Institute 2008a, 2008b; code to replicate the analysis using STATA's *mi* procedure is provided online). This software uses unweighted data in the imputation equation. Consistent with common econometric practice for complete-data analysis, and to avoid issues of 'uncongeniality' between imputation and analysis models, our analysis equation is also unweighted.

Our analysis equation uses composites of E_{t-2} and E_{t-1}. We consider three durations l of full-time employment spells in progress at the time of birth exposure D_l: 0, 1, and 2+ years. The reference category is $D_0 \equiv \{1\ if\ E_{t-1} = 0\ and\ 0\ if\ E_{t-1} = 1\}$ and therefore requires only information from $t - 1$. To code the alternate categories of duration of exactly 1 year, D_1, and duration of two or more years, D_2, information at both times $t-2$ and $t-1$ is required, since $D_1 \equiv \{1\ if\ E_{t-1}=1\ and\ E_{t-2}=0\}$ and $D_2 \equiv \{1\ if\ E_{t-1}=1\ and\ E_{t-2}=1\}$. The analysis model we estimate is then:

$$LOGIT[\Pr\{Y_t = 1|D_2, D_1, Z_{t-1}\}] = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 Z_{t-1} \qquad (2)$$

The expected efficiency gain in the estimation of β_3 when multiply imputing employment status at $t-2$, relative to estimation with the complete data only, is approximately equal to the fraction of observations with missing imputing employment status at $t-2$ (Little 1992, White and Carlin 2010, and see discussion in Rendall and Greulich 2014). The expected efficiency gains in the estimation of β_1 and β_2 are a priori unknown, but are of particular interest here. Because durations D_1 and D_2 are composites of employment status in times $t-2$ and $t-1$, they are constructed partially from observed data and partially from multiply imputed data. Therefore the reductions in $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$ may be substantial, even though less than the reductions in $Var(\hat{\beta}_3)$. For $Var(\hat{\beta}_3)$, the corresponding variable vector Z is entirely constructed from observed data, and therefore the reduction in variance is expected to approximate the fraction missing.

3. Results

The sample consists of person-years exposed to a first birth among partnered, parity-0 women (see Table 1). For only 200 person-years do we observe the woman's employment status two years before her calendar year of exposure to first birth. Of a total of 671 person-year observations, 323 were full-time employed in the year before birth exposure $(t-1)$ and did not have employment status observed in the year before that $(t-2)$ because they had not yet entered the panel. These are the left-censored spells. The 'fraction missing' is then 0.481 (323/671). Across all person-years, a weighted 74.5% of women were full-time employed the year immediately preceding exposure to first birth, and 70.8% were full-time employed two years before exposure.

Table 1: Descriptive statistics and numbers of observations, partnered parity-0 Polish women ages 18–39, 2005–2012

Descriptive statistics (person-years, weighted)			
	Mean	Standard deviation	Sample size
woman's age	28.7	4.4	671
woman's full-time employment in t–1 (proportion)	0.745	0.436	671
woman's full-time employment in t–2 (proportion)	0.708	0.456	200
partner's full-time employment in t–1 (proportion)	0.859	0.349	671
Number of observations (unweighted person-years)			
Left censored observations: full-time employed at t–1, not observed at t–2			323
Non-left-censored observations			348
All observations			671
Fraction missing full-time employment duration (left-censored)			0.481

Note: all observations have valid values of age and partner’s employment status at t-1 and of birth between t and t+1. All statistics are weighted.
Source: European Union Survey of Income and Living Conditions, Poland 2005–2012

Table 2: Logistic regressions of birth in year t, before and after imputing full-time employment status in t–2, partnered parity–0 Polish women ages 18 to 39, 2005–2012

	Model 1				Model 2			
	complete data, reduced specification[a]				complete data, full specification[b]			
	Estimate	Odds Ratio	Stand-ard Error	p-value	Estimate	Odds Ratio	Stand-ard Error	p-value
Intercept	–1.433		0.38	< 0.001	–1.413		0.48	0.003
Full-time-employed at t–1	0.658	1.93	0.26	0.011				
Full-time employed at t–1 and t–2					0.675	1.96	0.35	0.056
("duration 2+ years")								
Full-time employed at t–1 but not t–2					–0.112	0.89	0.67	0.867
(“duration 1 year”)								
age – 18	–0.102		0.03	< 0.001	–0.111		0.04	0.006
partner full-time employed at t–1	0.555		0.34	0.103	0.617		0.459	0.179
Sample	671				348			

a. Excludes employment status at time t–2.

b. Includes regressors constructed from employment status at time t–2.

c. Calculated by squaring the standard errors and taking the proportionate reduction in these variances about the parameter estimate from Model 2 to Model 3.

All regressions are unweighted.

Source: European Union Survey of Income and Living Conditions, Poland 2005–2012

Table 2: (Continued)

	Model 3				
	complete and multiply-imputed data, full specification[b]				
	Estimate	Odds Ratio	Stand-ard Error	p-value	Reduction in variance, Model 2 to Model 3[c]
Intercept	–1.407		0.38	< 0.001	0.365
Full-time-employed at t–1					
Full-time employed at t–1 and t–2	0.724	2.06	0.29	0.012	0.342
(“duration 2+ years”)					
Full-time employed at t–1 but not t–2	0.365	1.44	0.59	0.539	0.220
(”duration 1 year”)					
age – 18	–0.106		0.03	< 0.001	0.536
partner full-time employed at t–1	0.559		0.34	0.104	0.440
sample n	671				

a. Excludes employment status at time t–2.

b. Includes regressors constructed from employment status at time t–2.

c. Calculated by squaring the standard errors and taking the proportionate reduction in these variances about the parameter estimate from Model 2 to Model 3.

All regressions are unweighted.

Source: European Union Survey of Income and Living Conditions, Poland 2005–2012

Regression results are presented in Table 2. In Model 1, in which all 671 person-years are used, but for which the specification of employment status is reduced to one prior year, being full-time employed at $t - 1$ is associated with a 1.93 greater odds of giving birth. This result is consistent with Matysiak's (2009) finding using retrospective data, in which she also used employment status only in the year immediately before exposure. She estimated the model without partner's employment status among her predictors, explaining (p. 260) that partner data was missing for more than half the female sample. Our specification instead takes advantage of the partner employment-status variable, which is both better obtained from a panel survey than a retrospective survey and has strong justification in the theory and evidence on couple fertility (e.g., Vignoli, Drefahl, and De Santis 2012).

Model 2 distinguishes between 1 year only and 2+ years of full-time employment, and is estimated with the 348 person-years for which these durations are observed in the complete data. Having been full-time employed 2 or more years ('duration 2+') is associated with a 1.96 greater odds of giving birth compared to not having been full-time employed in the prior year ('duration 0'). This is statistically significant, however, only at the 0.10 level (p = 0.06). Having become full-time employed only in the most recent year ('duration 1') is not a statistically significant predictor of giving birth. These results are suggestive of duration of full-time employment being a critical factor in predicting a partnered woman's first birth. When restricted to using complete data, however, we are only able to include employment duration in the model at the cost of eliminating almost half of an already small sample, thereby rendering both employment-duration coefficients non-significant at conventional thresholds.

Our preferred model is Model 3, in which all 671 person-year observations are used, and with a specification of full-time employment that distinguishes 0, 1, and 2+ years' duration. This is the model made possible by multiply imputing values of the full-time-employed variable for the 323 person-years in which the woman was observed as full-time employed at time $t - 1$ and was not observed at time $t - 2$. Being full-time employed for 2 or more years is associated with a 2.06 greater odds of giving birth compared to not having been full-time employed in the prior year ('duration 0'), and the coefficient is now significant at the 0.05 level (p = .01). Being full-time employed only in the most recent year ('duration 1') is again not statistically significant. The proportionate reductions in variances about the coefficients for age (0.536) and partner's employment status (0.440) approximate the fraction missing (0.481). The proportionate reductions in variances are respectively 0.342 and 0.220 for the coefficients for full-time employed two or more years and for full-time employed only one year. These reductions are substantially less than the fraction missing, as expected, but are nevertheless quite large.

4. Conclusion

In short panels and in panels that sample from populations rather than from cohorts, such as the EU-SILC of the present study and the U.S. Survey of Income and Program Participation (SIPP, US Census Bureau 2014), left censoring is present for almost every individual. Supplementary histories collected retrospectively may be much less accurate than panel collection (Jacobs 2002; Kyyra and Wilke 2014), and not all characteristics of state or event histories will be covered; for example, parent-child co-residence. We proposed MI as a general solution to the problem of left censoring in demographic hazard modeling. As an example, we examined the gains that may be realized by multiply imputing a single additional year of employment status before the first wave of the panel. This was the maximum possible amount of imputation in the four-wave EU-SILC. Nevertheless, it allowed us to conduct more effectively a simple test of the hypothesis that women are more likely to begin childbearing after first obtaining stable employment (Santarelli 2011). Using conventional methods to conduct this test would have required using only half the number of person-year observations that we were able to use in our multiply imputed data analysis.

Substantively, we found that being full-time employed for two or more years was strongly predictive of a birth. Only in the analysis with the multiply imputed data, however, was the coefficient statistically significant at conventional levels ($p < .05$). The magnitude of variance reduction about this coefficient in the multiply imputed data was around one-third. To have obtained a variance reduction of this size represents a substantial payoff to having multiply imputed the source variable (employment status at $t-2$) for a large fraction of the person-year sample. We attribute the large variance reduction to the fact that for every observation at least some information was available on the length of the employment spell. Future work, however, might profitably investigate the different amounts of variance reduction that may be realized under different types and magnitudes of missing versus non-missing information in left-censored histories.

5. Acknowledgements

We are grateful to the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development for support under population research infrastructure grant R24HD041041, and to Yeats Ye for programming assistance. The data used in this study is from the European Commission, Eurostat, longitudinal EU SILC, years 2005–2012 (up to the October 2014 release). Eurostat has no responsibility for the results and conclusions of the authors.

References

Allison, P.D. (2001). *Missing Data*. Thousand Oaks: Sage Publications.

Eurostat (2011). 2008 Comparative EU final quality report. Version 3, July 2011.

Iacovou, M., Kaminska, O., and Levy, H. (2012). Using EU-SILC data for cross-national analysis: Strengths, problems and recommendations. Institute for Social and Economic Research Working Paper Series No. 2012-03, Essex University.

Jacobs, S.C. (2002). Reliability of recall and unemployment events using retrospective data. *Work, Employment, and Society* 16(3): 537–548. doi:10.1177/095001702762217489.

Johnson, D.R. and Young, R. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family* 73(5): 926–945. doi:10.1111/j.1741-3737.2011.00861.x.

Kyyra, T. and Wilke, R.A. (2014). On the reliability of retrospective employment information in European Household Panel data. *Empirical Economics* 46(4): 1473–1493. doi:10.1007/s00181-013-0718-1.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* 87(420): 1227–1237. doi:10.2307/2290664.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd Edition). Hoboken: Wiley. doi:10.1002/9781119013563.

Matysiak, A. (2009). Employment first, then childbearing: Women's strategy in post-socialist Poland. *Population Studies* 63(3): 253–276. doi:10.1080/00324720903151100.

Meng, X-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9(4): 538–573.

Moffitt, R.A. and Rendall, M.S. (1995). Cohort trends in the lifetime distribution of female family headship in the U.S., 1968–85. *Demography* 32(3): 407–424.

Özcan, B., Mayer, K.U., and Luedicke, J. (2010). The impact of unemployment on the transition to parenthood. *Demographic Research* 23(29): 807–846. doi:10.4054/DemRes.2010.23.29.

Raghunathan, T.E. and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association* 90(429): 54–63. doi:10.1080/01621459.1995.10476488.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27(1): 85–95.

Rendall, M.S. and Greulich, A. (2014). Multiple imputation for demographic hazard models with left-censored predictor variables. Maryland Population Research Center Working Paper PWP-MPRC-2014-011.

Santarelli, E. (2011). Economic resources and the first child in Italy: A focus on income and job stability. *Demographic Research* 25(9): 311–336. doi:10.4054/DemRes.2011.25.9.

SAS Institute (2008a). The MI Procedure. Chapter 54 SAS/STAT 9.2 User Guide, 2nd Ed.

SAS Institute (2008b). The MIANALYZE Procedure. Chapter 55 SAS/STAT 9.2 User Guide, 2nd Ed.

Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147–177. doi:10.1037/1082-989X.7.2.147.

US Census Bureau (2014). SIPP Introduction and History. http://www.census.gov/programs-surveys/sipp/about/sipp-introduction-history.html#.
[Accessed October 13, 2014]

Vignoli, D., Drefahl, S., and De Santis, G. (2012). Whose job instability affects the likelihood of becoming a parent in Italy? A tale of two partners. *Demographic Research* 26(2): 41–62. doi:10.4054/DemRes.2012.26.2.

White, I.R. and Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 29(28): 2920–2931. doi:10.1002/sim.3944.

Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review* 71(3): 581–592. doi:10.1111/j.1751-5823.2003.tb00213.x.