Revisiting the Political Life-Cycle Model: Later Maturation and Turnout Decline Among Young Adults

Kaat Smets
Royal Holloway, University of London
e-mail: kaat.smets@royalholloway.ac.uk

Published in European Political Science Review

Online Appendix C

Handling missing data raises both conceptual difficulties and computational challenges. The default way in which most statistical packages approach the missing values problem – through listwise deletion or complete case analyses – can yield bias, is inefficient, and is therefore considered unreliable. In general, two approaches to handling missing data are recommended in the literature: maximum likelihood (ML) and multiple imputation (MI) (Schafer and Graham, 2002; Raghunathan, 2004).

There are different types of missing data. Unit non-response occurs when the entire data collection procedure fails because respondents e.g. refuse to participate. Item non-response occurs when data are partially missing because a respondent e.g. did not answer all questions in the survey. Although part of the missing data problem in this paper is due to item non-response, the more pressing problem is that certain variables were not included in certain waves of the British Election Studies. Using listwise deletion would cause the sample size to vary considerably in the analyses presented. This renders comparison of results between the models difficult.

Table 1 below lists all the variables used in the analyses of this paper, the ratio of missing to valid answers, the percentage of missing values[1],and the main reason for the lack of data. If no reason is given, data are missing due to item non-response. The variable with the largest percentage of missing values is residential stability, which was missing in 27.5% of the cases.

[1]Ratio's and percentages are calculated based on the whole sample rather than the subset of young adults aged 35 years or less. The reason for this is that the average turnout of older citizens is part of the model and missing values thus need to be imputed for all respondents.

variable name	missing/valid value ratio	percentage of missing values	reason for missingness
turnout	0/33737	0%	
age	0/33737	0%	
left education	203/33737	0.6%	
married	29/33737	0%	
has child(ren)	6915/33737	20.5%	Not included in 1974 (feb), 1992
home ownership	2000/33737	5.9%	Not included in 1966
residential stability	9264/33737	27.5%	Not included in 1983, 1992; 1/2 sample in 2001
works	57/33737	0%	
gender	0/33737	0%	
educational level	897/33737	2.7%	
union membership	3014/33737	8.9%	Not included in 1966
pid strength	1656/33737	4.9%	
voted in previous elections	4690/33737	13.9%	Only posed to 1/6 of the sample in 2001
average turnout older voters	0/33737	0%	

Table 1: Missing values of modelled variables

The percentage of missing values in the sample used is quite high for certain variables. This calls for an imputation method with a high level of efficiency. Suppose x is a real value and $\hat{x}$ an estimated value. While treating missing data in a sample we want to make sure that the bias between estimated and the true values is small. Moreover, we want the variance and standard deviation of the estimated values to be small. Bias and variance are often combined into one measure called mean square error, which is the squared distance between the estimated and the real values over repeated samples: $(\hat{x}\text{-}x)^2$. The mean square error is equal to the squared bias plus the variance. Bias, variance, and the mean error describe the behaviour of an estimate. However, we also want to be confident about the measures of uncertainty that we report and estimate the true x with a probability of a certain predefined rate (Schafer and Graham, 2002, p. 149).

Multiple imputation (MI) is a method for handling missing data that solves the problem of uncertainty that many single imputation methods face. MI replaces each missing value by a list of $m > 1$ simulated values and as such produces m plausible alternative versions of the complete data set. Each of the m data sets is estimated in the same fashion by a complete data method. Estimates of parameters of interest are subsequently averaged to give a single estimate. Standard errors are computed according to the 'Rubin rules' (see below), allowing for between- and within-imputation components of variation in the parameter estimates.

MI does not need many rounds of estimation to reach a high level of efficiency. Rubin (1987) developed with the following way to calculate the efficiency of an estimate based on an m number of imputation (see equation 1):

$$eff = (1 + \lambda/m)^{-1} \tag{1}$$

where the efficiency is a function of the rate of the missing information (λ) and the number of imputations (m). For example, with 27.5% of missing information (as is the case with residential stability), m= 5 imputations will yield results that are $100/(1 + .055) = 94.8\%$ efficient. A rule of thumb for selection of the number of imputation rounds is that the confidence coefficient for the worst-case parameter (in this case residential stability) should be at least 95% (Royston, 2004, p. 239). This means that

in this particular case more than five rounds of imputation are desirable. Six rounds of imputation yield an efficiency of 95.6% for the residential stability variable. Therefore, m is set to six for the imputation procedure used to handle missing data for the analyses in this paper. The `ice` command in Stata is used to execute the multiple imputation process (see Royston, 2004, 2005a,b).

As mentioned above, multiple imputation creates a small number of data sets (in this case six), each of which has the missing values suitably imputed. The next step is to analyze each complete data set independently and summarize the results of these independent estimations. Coefficients are simply averaged. Summarizing the standard errors requires a bit more work (see equation 2 taken from Rubin (1987)):

$$s = \sqrt{\bar{u}_m + \frac{m+1}{m} b_m} \tag{2}$$

where $\bar{u}_m$ is the mean of the standard error's, and b_m is the variance of the estimates across the imputations. The `micombine` command in Stata combines the estimates from the m analyses using Rubin's rules (Royston, 2004, 2005a,b).

References

Raghunathan, T. E. (2004). What do we do with missing data? some options for analysis of incomplete data. *Annual Review of Public Health 25*, 99–117.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal 4*(3), 227–241.

Royston, P. (2005a). Multiple imputation of missing values: Update. *The Stata Journal 5*(2), 188–201.

Royston, P. (2005b). Multiple imputation of missing values: Update of ice. *The Stata Journal 5*(4), 527–536.

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

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods 7*(2), 147–177.