Imputation of Public Sewer Use in the 2017 American Housing Survey

Updated November 2018

Background

The American Housing Survey (AHS) collects data on a number of housing quality items every two years. Public sewer use is a core part of this questionnaire. Survey respondents are asked "Is your home connected to a public sewer?" and are given the option to respond with a yes, no, don't know, or refuse to answer.

In years prior to 2015, Census imputed the don't know and refused values using an assumed rate of public sewer use taking into account a unit's urban/rural status. In 2015, Census removed this imputation in 2015, leaving these non-responses as missing values on the public use files. This document describes the new imputation methodology first used in 2017.

Imputation Methodology

Census investigated the potential usefulness of variables as predictors in an imputation model and found structure type and housing-unit density to be useful predictors of public sewer use. Structure type is collected in the AHS; housing-unit density was obtained from the decennial census.

Missing data in public sewer use were imputed using a stochastic regression imputation approach. This procedure results in more accurate imputed data than most other traditional approaches such as hot decks and regression imputation[1].

Stochastic regression imputation uses regression equations to predict incomplete variables (public sewer) from complete variables (structure type and housing-unit density). By itself this can result in imputed values that are biased[2]. In particular, it will overestimate the strength of the relationship between the variable being imputed and the predictors. Additionally, it will reduce variances and covariances. Stochastic regression imputation addresses these issues by adjusting each predicted value with a normally distributed residual term. Adding a residual to the predicted values restores the lost variability to the data and largely addresses the aforementioned biases.

Results

This model was developed using 2015 AHS data and implemented in 2017 AHS data. It was developed and implemented using multiple survey years of data independently in order to increase the likelihood that the models would be generalizable across years. Models that are

[1] Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.
[2] Enders, 2010.

developed and implemented in a single survey year may not be useful in other years given that the relationship between variables can change over time.

The previously described imputation methodology resulted in a distribution of imputed values similar to the non-imputed values. There are 705 cases with missing public sewer data that were eligible to be imputed. Among cases that have a valid value for public sewer prior to imputation (i.e. non-imputed values), 85.88% of cases are connected to a public sewer and 14.12% of cases are not connected to one. Similarly, 77.45% of imputed cases are connected while 22.55% are not[3]. Including all cases, 85.82% of cases are connected to a public sewer while 14.18% are not. This suggests that our imputation methodology did not significantly affect the distribution of public sewer data.

Conclusion

This model-based imputation addresses the missing data in public sewer use using a methodology designed to ensure accurate estimates. The implementation of this approach was preferable to alternatives such as hot-deck imputation. A key advantage of model-based imputation is that it readily allows for the inclusion of continuous predictors. In order to include continuous predictors in a hot deck they need to be transformed into categorical variables. This results in the loss of information in the predictor variables[4-5]. Consequently, hot-deck imputation was not ideal since there were useful continuous predictors. In particular, the stochastic regression imputation approach described in this paper allowed us to reliably and accurately impute public sewer. This approach avoided the biases that can occur in other forms of imputation.

[3] These are the percentages immediately after imputation. Some values were later edited, slightly altering these percentages.
[4] Enders, 2010.
[5] Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology*, *47*(3), 537-560.