MISSING DATA IN AN ENVIRONMENTAL EXPOSURE STUDY: IMPUTATION TO IMPROVE SURVEY ESTIMATION

Olson K[1], Sinibaldi J[1], Lepkowski J[1], Ward B[1], Ladronka K[1], Towey T[2], Wright D[2], Gillespie BW[3]

[1]Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, Michigan 48104; [2]Department of Civil and Environmental Engineering, University of Michigan College of Engineering, 2340 G.G. Brown Building, Ann Arbor, MI 48109; [3]Department of Biostatistics, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109

Keywords: Humans, Survey interviews, Dietary intake, Occupation, Fish, Game consumption, Activities, Quality Control

Introduction

The University of Michigan Dioxin Exposure Study (UMDES) was undertaken in response to concerns among the population of Midland and Saginaw counties in Michigan that dioxin-like compounds from the Dow Chemical Company facilities in Midland have contaminated soils in the Tittabawassee River flood plain and areas of the city of Midland. The UMDES was designed to answer the following questions: 1) Are dioxin levels in serum increased among people who live in the Tittabawassee River flood plain compared to people who live elsewhere in Midland and Saginaw counties in elsewhere in the state of Michigan? 2) What factors explain the variation is serum dioxin levels among the entire population?

In a study with such important issues, the UMDES study team decided that rigorous and up-to-date methodology was required in all phases of the study. Methods used in data collection and analysis as well as the study findings are reported elsewhere.[1,2,3,4,5,6] This paper describes the methods used to handle item missing values.

Survey data collection inevitably yields data which are missing for study units (in this case, persons ages 18 or older living in their residence for the last five years selected into the study sample) or for items for a responding person who did not know an answer or refused to provide a response. Many analysts when faced with such missing data in a survey choose to ignore it. This strategy has several limitations that can be overcome in part by survey weights to compensate for unit nonresponse (see Lepkowski et al., for the UMDES weighting methods)[7] or assigning values to replace item missing values, or imputations. The purpose of this paper is to report methods used to impute for item missing values in the UMDES.

Materials and Methods

Three problems arise if analysts choose to ignore missing values in an analysis: 1) sample sizes are decreased, and estimates from survey data are less precise; 2) bias can be introduced into survey estimates due to missing data from individuals not responding who are different from those who do respond; and 3) substantial sample size losses can occur for analyses using several variables simultaneously, such as in multiple linear regression models.

Missing values are often replaced with imputed values, providing analysts with complete data. Imputation methods are varied, although most are a form of regression imputation.[8] For example, when an analyst chooses to ignore item missing values in estimating a simple statistic such as the mean serum TEQ among responding eligible persons in the UMDES, a strong “missing completely at random” assumption is being made that the missing values have been selected at random from among all respondents who gave serum. Further, under this assumption, the mean among respondents is implicitly being imputed for each missing value in estimates of mean serum TEQ. In other words, the analyst implicitly imputes the overall mean for each missing value when they ‘ignore missing values.’.

Obviously the assumption of missing completely at random can be weakened, and missing values can be replaced by a more plausible substitute for the unknown correct value. For example, one could assume that subgroups exist for which the missing values are similar to the observed values within the subgroup. In the UMDES, serum TEQ values are expected to vary by age, with younger persons having lower values. Thus, a weaker assumption is that serum TEQ values are "missing at random" within age groups. The mean serum TEQ value in an age group could then be imputed for the missing serum values in the age group. Cell mean imputation is an improvement on the implicit mean value imputation employed under the 'ignore the missing value' strategy.

The assumption required for imputation can be further weakened by employing better predictive models, such as a linear regression model. Consider for serum TEQ a model with *p* total predictors for the respondents providing serum samples: $y_{ri} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{rji} + \varepsilon_{ri}$. Here, y_{ri} denotes the serum TEQ value for the i^th responding person, and x_{rji} the value of various predictors (e.g., age, location of residence, body mass index, consumption of fish from contaminated rivers, etc.) of serum TEQ values for that responding person. The estimated values of the coefficients b_j can be employed to obtain a predicted value for the i^th person with a missing value: $\hat{y}_{mi} = b_0 + \sum_{j=1}^{p} b_j x_{mji} + \hat{\varepsilon}_{mi}$, where a residual $\hat{\varepsilon}_{mi}$ could be randomly generated from one of several distributions (e.g., a Normal residual with mean zero and variance equal to the variance of the residuals of the non-missing cases).

The regression imputation has limitations. If the outcome variable is not continuous, linear regression may not be appropriate; other model forms such as logistic, Poisson, or multinomial logistic can be used. Without randomly generated residuals, individuals with missing serum values and the exact same values of the predictors x_{mi}, will have the same predicted value. That is, the randomly generated residual avoids 'spikes' of repeated values in the serum TEQ values. Even with randomly generated residuals, though, the regression imputation fails to preserve underlying covariance structures among the imputed variables. For example, suppose that serum values for two different congeners, 2,3,7,8-TCDD and 1,2,3,7,8-PentaCDF were imputed, and a separate regression model generated imputed values for each. If the regression models had different predictor sets, there would be a tendency to generate predicted values with lower correlations between TCDD and the furan than among the values for the respondents.

A method of sequential regressions[9] can be employed to preserve the underlying covariance structure for the entire dataset. Suppose the set of variables are divided into two groups, a set of *q* variables with no missing values (say, for example, age, location of residence, gender) and a set of *p* with missing values to be imputed (say, food consumption variables and serum values for 29 congeners). The approach is as follows: Using regression imputation, obtain predicted values for the first of the variables in the set of the *p* to be imputed using a suitable form of regression imputation (linear, logistic, etc.). Generate the predicted values for the first imputed variable as a function of the full set of variables that had no missing values, say $\hat{y}_1 = f\left(x_1, x_2, ..., x_q\right)$. For the second variable to be imputed, obtain predicted values from a regression prediction, using a suitable model form, with the full set of *q* variables without missing values *and* the now completed variable, $\hat{y}_1$ as part of the predictor set, or $\hat{y}_2 = f\left(x_1, x_2, ..., x_q, \hat{y}_1\right)$. Continuing the process, obtain predicted values from regression imputation for the third variable as $\hat{y}_3 = f\left(x_1, x_2, ..., x_q, \hat{y}_1, \hat{y}_2\right)$. Repeat this process for all *p* variables requiring imputed values.

This cycle of regression imputations may have an underlying order effect. To eliminate any order effect, repeat the regression imputations for each of the *p* variables, but employing all $q + p - 1$ variables as predictors. For example, re-impute the first variable needing imputed values using the *q* variables without missing values again plus all the remaining p – 1 variables that received imputed values in the first cycle: $\hat{y}_1 = f\left(x_1, x_2, ..., x_q, \hat{y}_2, \hat{y}_3, ..., \hat{y}_p\right)$. Re-impute all p variables with missing data in this way, repeating this cycle several times. Empirical investigations indicate that five complete cycles are satisfactory to minimize any order effect of the original imputation cycle.[9]

This sequential regression procedure has been automated in the IVEware software system (available from the University of Michigan).[9] IVEware was used to impute item missing values in the UMDES. Nearly all variables in the data set were imputed using the sequential regression imputation procedure.

Implementation involves a number of complexities. Forms of regression must be specified for all variables to be imputed. Restrictions and constraints must be specified to avoid imputing values for a variable that are inappropriate (e.g., imputing breast feeding practices for men) or implausible (e.g., doing an activity for a total number of years greater than the individual's age).

Low frequency and restricted variables provide limited data for estimating the regression models. For instance, in the UMDES the number of years employed in the chemical industry was answered by only a few subjects who were ever employed in the industry. It was not possible to estimate the coefficients because of small sample size.

Complex date variables may be extremely difficult to impute. For example, the UMDES asked for years in a person's life when they ate fish caught in the Tittabawassee River. If the person did not remember the dates, the years were missing. Imputation must preserve consistency, and three variables must be imputed sequentially: first whether the person ever ate fish caught from the Tittabawassee River, then the beginning year, and the ending year. The beginning year must be no earlier than the person's birth year, and no later than the year of interview. The ending year must be no earlier than the year after the beginning year and no later than the year of interview. A large number of such variable sequences in the UMDES data had to be programmed. Further, some date sequences covered rare events, such that there were very few cases from which regression coefficients could be estimated.

Results

Table 1 shows the distribution of item missing values for several UMDES variables. The frequency of item missing values was, for the most part, quite low.

Table 1. Number and percent missing for five variables in the UMDES.

Variable	No. reported values	No. missing values	Percent missing
Body mass index			
Years eating fish caught from Tittabawassee River			
Game meat meals eaten in last five years			
Serum TEQ (ppt)			

Table 2 shows the results of imputation for these same variables. The distribution of imputed values is not necessarily the same as the distribution of the original reported values. This is expected if the values of predictors for cases being imputed differ from those for those with reported values. Thus, one should expect the imputed values to differ from the reported values.

Table 2. Number and percent missing for five variables in the UMDES.

Variable	Mean, reported values	Mean, imputed values	Mean, all values
Body mass index			
Years eating fish caught from Tittabawassee River			

Game meat meals eaten in last five years
Serum TEQ (ppt)

Finally, Table 3 compares the coefficients for an important model in the UMDES, the regression of serum TEQ on age and body mass index. Two coefficients are given for each predictor: that obtained from cases with no missing values on serum TEQ or the predictors, and that obtained after imputation replaced item missing values for the variables in the model. In this case, there are few differences between the sets of estimated coefficients, indicating that the imputation had little effect on the final estimated model.

Table 3. Regression coefficients from reported values only and from reported and imputed values in UMDES model for serum TEQ

Variable	Reported values	All values
Age		
Age, squared		
Body mass index		

Discussion

Imputation is not a widely used technique in environmental exposure studies. Yet environmental exposure studies have item missing values. Analysts ignore item missing values, and thereby implicitly impute for item missing values under a strong assumption of completely missing at random. In the UMDES, item missing values were replaced by imputed values from a sequential regression imputation procedure. These imputed values allowed the analyst to estimate models from a complete data set, and improve the bias and variance properties of the subsequent estimates of means, proportions, and regression coefficients.

References

[1] Franzblau, A, Garabrant, D, Adriaens, P, Gillespie, B, Lepkowski, J, Olson, K, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006 (forthcoming).

[2] Olson, K, Garabrant, D, Franzblau, A, Adriaens, P, Gillespie, B, Lepkowski, J, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006 (forthcoming).

[3] Adriaens, P, Garabrant, D, Franzblau, A, Gillespie, B, Lepkowski, J, Olson, K, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006 (forthcoming).

[4] Zwica, L, Garabrant, D, Franzblau, A, Adriaens, P, Gillespie, B, Lepkowski, J, Olson, K, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D. *Organohalogen Comp* 2006 (forthcoming).

[5] Hedgeman, E, Garabrant, D, Franzblau, A, Adriaens, P, Gillespie, B, Lepkowski, J, Olson, K, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006 (forthcoming).

[6] Garabrant, D, Franzblau, A, Adriaens, P, Gillespie, B, Lepkowski, J, Olson, K, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006 (forthcoming).

[7] Kalton, G, Kasprzyk, D. *Survey Methodology* 1986, 1 - 16.

[8] Raghunathan, TE, Lepkowski, J, Van Hoewyk, J, Solenberger, P. *Survey Methodology* 2001, 85-96.

[9] www.isr.umich.edu/src/smp/iveware/