Unit 2 Case Study – Multiple Imputations

**Armand Post, Andrew Abbott, Mustafa Sakarwala**

*MSDS 7333 – Quantifying the World, 1/23/2018*

**ABSTRACT**

The multiple imputation was developed as a general method for inference with missing data. Instead of replacing the missing observation with a single value, multiple imputation method replaces each missing value with multiple plausible values. This study reviews multiple imputation as an analytic strategy for missing data and applies statistical methods to impute missing data on a dataset containing data related to the fuel economy of cars.

**Introduction**

Broadly, the term missing data means that we are missing some type of information about the phenomena in which we are interested. When analyzing data, in all fields of study, missing data is a common problem. Since, for any data collection process, there are so many things that could go wrong that missing values are all too likely. There are a variety of reasons data would be missing. We classify those into three broad categories related to (1) the study participants (2) the study design, and (3) the interaction of the participants and the study design. As the old saying goes, the only certainties are death and taxes. We would like to add one more to that list: missing data. When analyzing a dataset, missing values can decrease the power of the analysis. Some methods cannot be used at all with data not complete and missing.

A common approach for addressing missing data, wherein a single value is imputed for each missing observation, is known as single imputation. Single imputation methods have been the subject of increasing criticism with respect to their tendency to underestimate standard errors, overstate statistical significance, and introduce bias. However, if the proportion of missing values is small (less than 5%), then a simple imputation method may be considered to be accurate.

An alternative method for addressing incomplete data is multiple imputations (MI). Through this method, we can replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiple imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. MI does not attempt to estimate each missing value through simulated values but rather to represent a random sample of the missing values. This process results in the valid statistical inferences that more accurately reflect the uncertainty due to missing values.

**LITERATURE REVIEW**

As part of our research, a literature review was performed. Bergland’s paper from the 2010 SAS Global Forum provided insight into the use of sophisticated imputation. Statistically sophisticated imputation provides advantages over simple methods such as inserting mean values. Simpler methods do not account for introduced variability and distort variable distributions. Also, if missing data is monotone, analysts will have more flexibility when imputing. She also explained that continuous and categorical variables require different imputation approaches (Bergland, 2010).

Tom Rosenström of the University of Helsinki’s Lecture Notes: Some Core Ideas of Imputation for Nonresponse in Surveys also provided some insight into the use of imputation. Rosenström showed how excluding missing records in a sample dataset resulted in a small to medium relationship between two parameters while imputing showed a large correlation. Regression slopes were also shown that compared numerous methods. These showed results in line with Bergland’s assertions that more sophisticated methods are needed when dealing with missing data. In Rosenström’s words, “Imputation methods that introduce biased estimates or their standard errors, such as mean imputation and listwise deletion, are called improper imputations. Methods that produce unbiased estimates and also properly handle their uncertainty estimation are called proper imputations.” Some discussion of how classical statistical method vs Bayesian methods was also found in the Lecture Notes. Since Bayesian statisticians treat both model parameters and missing / unobserved data the same, as random variables, Bayesian methods are easily extended for missing data modeling (Rosenström, 2014).

**BACKGROUND**

In our case study, a dataset (Car Miles per Gallon) from a study of fuel economy of cars is used to demonstrate methods for multiple imputation and analysis of measured data. The dataset analyzed in this study consist of mileage data from 38 cars as measured in 2005. The variables contained in this dataset include are as shown in Table 1.

|  |  |
| --- | --- |
| Description of Variables | |
| Variable | Description |
| Auto | Make and model of car |
| MPG | Estimated miles per gallon |
| CYLINDERS | Number of cylinders in engine |
| SIZE | Engine displacement (larger number = bigger engine) |
| HP | Horsepower |
| WEIGHT | Weight of car |
| ACCEL | Acceleration |
| ENG\_TYPE  Table 1. Names and descriptions of the variables contained in the dataset | Engine type |

This data also contains missing values. The intended analysis focuses on results of the linear regression on this dataset before and after imputing the missing values and determine if multiple imputations had a meaningful impact on the analysis. This will allow us to explore the impact of multiple imputations. To perform the analysis in this case study, we will use SAS version 9.4.

Multiple imputations are the main analysis method that will be examined in this case study. Essentially, missing values are estimated using random samples of values that are present in other records. When the desired number of imputed data sets is created using these estimated values, these datasets can then be combined to form one dataset with no missing values.

The MI procedure (PROC MI) creates multiple imputed data sets for incomplete multivariate data. It uses methods that incorporate appropriate variability across the imputations. The method of choice depends on the patterns of missing data. Once the complete data sets are analyzed using standard SAS procedures, the new MIANALYZE procedure can be employed to generate valid statistical inferences about these parameters by combining results from the analyses.

**METHODS**

The steps used for this study are:

1. Explore the variable distribution and missing data patterns.

Before determining a method of imputation, we first want to explore the data variable distribution, investigate the missing values and determine if there is a pattern in the missing data for missing values. We will use PROC MI in SAS with the nimpute=0 option to examine the missing data pattern. From Table 2, the missing data has no pattern of missingness and cannot be sorted in a way that when a missing value is observed, all other values after that are also. Hence, we conclude that the data is arbitrarily missing or non-monotone. Since the dataset is classified as non-monotone MCMC is the recommended method of imputation.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Missing Data Patterns | | | | | | | | | |
| Group | MPG | CYLINDERS | SIZE | HP | WEIGHT | ACCEL | ENG\_TYPE | Freq | Percent |
| 1 | X | X | X | X | X | X | X | 18 | 47.37 |
| 2 | X | X | X | X | X | X | . | 2 | 5.26 |
| 3 | X | X | X | X | X | . | X | 1 | 2.63 |
| 4 | X | X | X | X | X | . | . | 1 | 2.63 |
| 5 | X | X | X | X | . | X | X | 3 | 7.89 |
| 6 | X | X | X | X | . | . | X | 1 | 2.63 |
| 7 | X | X | X | . | X | X | X | 5 | 13.16 |
| 8 | X | X | . | X | X | X | X | 2 | 5.26 |
| 9 | X | X | . | X | . | X | X | 1 | 2.63 |
| 10 | X | . | X | X | X | X | X | 2 | 5.26 |
| 11 | X | . | X | X | X | . | X | 1 | 2.63 |
| 12 | X | . | X | X | . | X | X | 1 | 2.63 |

Table 2. Missing data pattern analysis

1. Then we perform linear regression on the original dataset. The response variable is MPG. The explanatory variables used are CYLINDERS, SIZE, HP, WEIGHT, ACCEL, and ENG\_TYPE.
2. Create multiple imputed datasets to replace missing values using PROC MI. We can use Markov Chain Monte Carlo (MCMC) which is a collection of methods for simulating random draws from non-standard distributions. MCMC is used to create a small number of independent draws of the missing data from a predictive distribution, and these draws are then used for multiple-imputation inference. MCMC method is used to create five imputed datasets. A seed of 35399 is used to enable the analysis to be reproduced.
3. Use of PROC MIANALYZE for analysis of the output from the analysis step including accounting for the variability introduced during the imputation step.
4. We can then compare and summarize these imputed results with the non-imputed results and determine if multiple imputations had a noticeable effect on the outcomes.

**RESULTS**

1. Variable distribution and missing data pattern:

With our dataset identified as non-monotone from Table 2, it is recommended that we use Markov Chain Monte Carlo (MCMC) for PROC MI. Also, we see that only 47.37% of the data is complete and without missing data. Missing data was found in 18 of the 38 observations.

1. Variable distribution and missing data pattern:

Using list-wise deletion on the dataset prior to running the regression function results in only 20 of the total 38 observations present in the dataset, as seen in Table 3. There are 18 observations, representing 47.37% of the entire dataset, which are excluded from the analysis. This is an issue as the number of observations in the full dataset is already very small.

Table 3. Non-Imputed Regression, Read/Used Observations

|  |  |
| --- | --- |
| Number of Observations Read | 38 |
| Number of Observations Used | 18 |
| Number of Observations with Missing Values | 20 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
| Intercept | 1 | 70.14772 | 8.03838 | 8.73 | <.0001 |
| CYLINDERS | 1 | -3.33403 | 1.56072 | -2.14 | 0.056 |
| SIZE | 1 | 0.0228 | 0.03207 | 0.71 | 0.4918 |
| HP | 1 | -0.19546 | 0.08065 | -2.42 | 0.0338 |
| WEIGHT | 1 | -0.30623 | 5.13263 | -0.06 | 0.9535 |
| ACCEL | 1 | -0.78199 | 0.58264 | -1.34 | 0.2066 |
| ENG\_TYPE | 1 | 6.5988 | 3.59008 | 1.84 | 0.0932 |

Table 4. Parameter Estimates for Non-Imputed Regression

Table 4 above shows the parameter estimates of regression using list-wise deletion. These values serve as the original parameter estimate values for comparison with the imputed results. From Table 4, we can see that both the CYLINDERS and HP variables are close to the alpha threshold of 0.05, meaning they have statistical significance when predicting MPG.

1. Create multiple imputed datasets to replace missing values:

Next, we executed multiple imputation processes using Markov Chain Monte Carlo (MCMC) method and then linear regression on each of the 5 imputed datasets. The results of the regressions were obtained via PROC MIANALYZE. The results of the comparisons of the estimates from the linear regressions can be found in Table 5 below.

Table 5. Parameter Estimates for Multiple Imputation Regression by Imputation

c

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Original Parameter Estimate | Imputation 1 Parameter Estimate | Imputation 2 Parameter Estimate | Imputation 3 Parameter Estimate | Imputation 4 Parameter Estimate | Imputation 5 Parameter Estimate |
| Intercept | 70.14772 | 71.07423 | 72.39109 | 68.6843 | 68.5524 | 67.01166 |
| CYLINDERS | -3.33403 | -3.03737 | -2.9346 | -2.93177 | -2.85923 | -2.69887 |
| SIZE | 0.0228 | 0.02391 | 0.02052 | 0.02557 | 0.04358 | 0.04106 |
| HP | -0.19546 | -0.15919 | -0.19765 | -0.14873 | -0.15208 | -0.13697 |
| WEIGHT | -0.30623 | -2.03889 | -0.26845 | -2.97682 | -4.65964 | -6.12984 |
| ACCEL | -0.78199 | -0.91547 | -1.07191 | -0.69925 | -0.57066 | -0.35255 |
| ENG\_TYPE | 6.5988 | 5.74751 | 6.22872 | 5.80842 | 5.19245 | 6.29941 |

Table 5: Parameter Estimated for Multiple Imputation Regression

1. Summarize and compare imputed data:

We find from comparisons in Table 6 the parameter estimates are very similar between MCMC imputation and original list-deletion methods. However, for all the variables, the standard errors are smaller for the combined estimates of imputed datasets. The imputed datasets had double the number of observations than the list-wise deletion dataset. This resulted in higher power for the analysis on the imputed data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Linear Regression Parameter Estimate | | | | |
| Variable | Original Parameter Estimate | Original Std Error | Combined Parameter Estimate | Combined Std  Error |
| Intercept | 70.14772 | 8.03838 | 69.542738 | 4.676262 |
| CYLINDERS | -3.33403 | 1.56072 | -2.892369 | 0.766712 |
| SIZE | 0.0228 | 0.03207 | 0.030931 | 0.021663 |
| HP | -0.19546 | 0.08065 | -0.158924 | 0.046085 |
| WEIGHT | -0.30623 | 5.13263 | -3.214728 | 3.740139 |
| ACCEL | -0.78199 | 0.58264 | -0.721966 | 0.409842 |
| ENG\_TYPE | 6.5988 | 3.59008 | 5.855301 | 1.579569 |

Table 6. Comparison of regression model parameter estimates using list-wise deletion and multiple imputation methods.

**CONCLUSION**

The original dataset for data related to the fuel economy of cars had 18 observations with missing data. When we applied the regularly used list-wise deletion regression to model and predict MPG, we were excluding 47.37% of the data. This introduced an unwanted bias into our analysis. To avoid such a bias, we applied MCMC multiple imputations so we could use all records instead of just the subset.

The objective of this case study was to demonstrate multiple imputations would produce better results than list-wise deletion when performing linear regression on the given fuel economy. The linear regression model for MPG prediction produced parameter estimates of greater statistical significance with smaller standard errors when imputation was performed compared to the list-wise deletion method with lesser observations.

To extend the effort, future work would be to refine the linear regression model by fine-tuning the exploratory variables used. This would be done by evaluating the Variable Inflation Factor (VIF) values and p-values for each variable and then looking at the adjusted R2 values. Also, we would use libraries available in R and/or Python which have more options to improve the results than are available in SAS.

**REFERENCES**

Patricia A. Berglund, Institute for Social Research-University of Michigan, Ann Arbor, Michigan.  
*An Introduction to Multiple Imputation of Complex Sample Data using SAS (Version 9.02).*  
<http://support.sas.com/resources/papers/proceedings10/265-2010.pdf>

Tom Rosenström, University of Helsinki

*Lecture Notes: Some Core Ideas of Imputation of Nonresponse Surveys*

http://www.helsinki.fi/~rosenstr/papers/ImputationNotes.pdf

**Appendix A**

/\*Datalines are used to make the work more easily reproducible \*/

DATA CarMPG;

INPUT Auto $23. MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

DATALINES;

Buick Estate Wagon 16.9 8 350 155 4.36 14.9 1

Ford Country Sq. Wagon 15.5 8 351 . 4.054 14.3 1

Chevy Malibu Wagon 19.2 8 267 125 3.605 15 1

Chrys Lebaron Wagon 18.5 8 360 150 3.94 13 1

Chevette 30 4 98 68 2.155 16.5 0

Toyota Corona 27.5 4 134 95 2.56 14.2 0

Datsun 510 27.2 4 119 97 2.3 14.7 0

Dodge Omni 30.9 4 105 75 2.23 14.5 .

Audi 5000 20.3 5 131 . 2.83 15.9 0

Volvo 240 GL 17 6 163 125 3.14 13.6 0

Saab 99 GLE 21.6 . 121 115 2.795 15.7 0

Peugeot 694 SL 16.2 6 . 133 3.41 15.8 0

Buick Century Spec. 20.6 . 231 105 3.38 15.8 0

Mercury Zephyr 20.8 6 200 85 . 16.7 0

Dodge Aspen 18.6 6 225 110 3.62 18.7 0

AMC Concord D/L 18.1 . 258 120 3.41 . 0

Chevy Caprice Classic 17 . 305 130 . 15.4 1

Ford LTD 17.6 8 302 129 3.725 . .

Mercury Grand Marquis 16.5 8 351 138 3.955 13.2 1

Dodge St Regis 18.2 8 318 135 3.83 . 1

Ford Mustang 4 26.5 4 140 . 2.585 14.4 0

Ford Mustang Ghia 21.9 6 171 . 2.91 16.6 1

Mazda GLC 34.1 4 86 65 . 15.2 0

Dodge Colt 35.1 4 98 80 1.915 14.4 0

AMC Spirit 27.4 4 121 . 2.67 15 0

VW Scirocco 31.5 4 89 71 1.99 14.9 0

Honda Accord 29.5 4 98 68 . 16.6 0

Buick Skylark 28.4 4 151 90 2.67 16 0

Chevy Citation 28.8 6 173 115 2.595 11.3 1

Olds Omega 26.8 6 173 115 2.7 12.9 1

Pontiac Phoenix 33.5 4 151 90 2.556 13.2 0

Plymouth Horizon 34.2 4 105 70 2.2 13.2 0

Datsun 210 31.8 4 85 65 2.02 19.2 .

Fiat Strada 37.3 4 91 69 2.13 14.7 0

VW Dasher 30.5 4 . 78 . 14.1 0

Datsun 810 22 6 . 97 2.815 14.5 0

BMW 320i 21.5 4 121 110 . . 0

VW Rabbit 31.9 4 89 71 1.925 14 0

;

RUN;

TITLE 'Predicting MPG';

PROC REG DATA = CarMPG;

MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

RUN;

QUIT;

ODS SELECT MISSPATTERN;

PROC MI DATA=CarMPG NIMPUTE=0;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

RUN;

PROC MI DATA = CarMPG

OUT = MIOUT seed = 35399;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

RUN;

PROC REG data = miout outest = outreg covout ;

Model MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

by \_Imputation\_;

RUN;

PROC MIANALYZE data = outreg ;

MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE Intercept;

RUN;