Case Study Unit 2 – Multiple Imputation

Cory Adams, Chris Boomhower, Alexandra Fisher, Alex Frye

MSDS 7333, September 12, 2017

ABSTRACT

Missing data is a known and all-too-common problem in studies involving the collection of data. Missing data can result in biased statistical analyses or even negatively impact a study's statistical power. This case study was undertaken in order to highlight and address the issue of missing data by comparing analyses using multiple imputation and listwise deletion. The dataset used contains a total of 38 records, each representing a unique car make and model, with observations recorded for 7 variables. The dataset had 25 total missing observations. Using SAS statistical software, the dataset's pattern of missing data was found to be arbitrary, or non-monotone. The recommended method of imputation based on the non-monotone patterns of missing data was, therefore, Markov Chain Monte Carlo (MCMC) full-data imputation. Multiple imputation was used to create five imputed datasets in which missing data values were replaced with imputed values through the use of selected regression techniques. The imputed datasets were then fitted using a linear regression model and the results were combined for comparison to a regression model fitted with listwise deletion data. Next steps involve case study replication using R as the statistical software.

INTRODUCTION

As various types of data are gathered into datasets for statistical analysis, it is common to find that recorded data is either incomplete or missing variables. Missing data can range from variables without observations to specific data questions without any known answers. Causes include human error in data collection, system malfunctions, corruption of data, or a number of other unintended origins. Statistical procedures conducted using statistical software, such as SAS or R, make it common practice to automatically eliminate instances of missing data. For example, regression analysis in SAS, as will be observed in this paper, handles the issue of missing data through listwise deletion, excluding records containing missing data from calculated estimates. The dataset of interest in this case study is the car fuel efficiency dataset which contains missing data for a number of variables. The objective of this paper is to highlight and address the use of a dataset with missing data through implementation of multiple imputation to produce unbiased estimates and optimize statistical power. Results will be compared to the less ideal solution of the listwise deletion method.

LITERATURE REVIEW

Multiple imputation concepts and theory were reviewed via Alan Elliot's (Southern Methodist University) asynchronous class lectures and Tom Rosenström's (University of Helsinki) lecture notes on imputation methodologies[1]. SAS specific methods were adopted via Alan Elliot's programming approaches outlined in his book on SAS programming, SAS Essentials[2]. Between these three sources, multiple imputation general concepts and SAS code were derived to perform this case study.

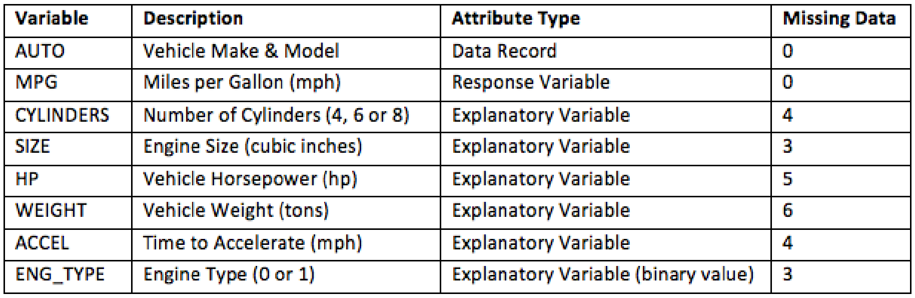
[1] Rosenström, T. (2014), Lecture Notes: Some Core Ideas of Imputation for Nonresponse in Surveys, University of Helsinki

[2] Elliot, A. (2015), SAS Essentials: A Guide to Mastering SAS, 2nd Edition, Wiley

BACKGROUND

The car dataset contains 38 total observations with 7 variables. Each record contains a different car make and model. Variable names, descriptions, and number of missing observations are included in Table 1.

Table 1 – Dataset Summary



Variable	Description	Attribute Type	Missing Data
AUTO	Vehicle Make & Model	Data Record	0
MPG	Miles per Gallon (mph)	Response Variable	0
CYLINDERS	Number of Cylinders (4, 6 or 8)	Explanatory Variable	4
SIZE	Engine Size (cubic inches)	Explanatory Variable	3
HP	Vehicle Horsepower (hp)	Explanatory Variable	5
WEIGHT	Vehicle Weight (tons)	Explanatory Variable	6
ACCEL	Time to Accelerate (mph)	Explanatory Variable	4
ENG_TYPE	Engine Type (0 or 1)	Explanatory Variable (binary value)	3

A scatterplot matrix with histograms shown below in Figure 1 reveals a negative correlation among four of the explanatory variables—number of cylinders, engine size, horsepower, and weight—and a vehicle's miles per gallon (MPG). Only the top row of the full correlation matrix is provided to show correlation among covariates and MPG. The scatterplots for acceleration and engine type show low correlations with MPG and, therefore, will not be used as variables of interest in the regression of MPG.

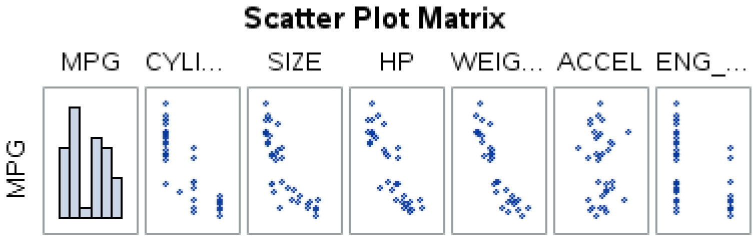


Figure 1 – MPG vs. Independent Variable Scatter Plot Matrix

METHODS

The steps used for this analysis were: 1) Analyze missing value patterns; 2) create multiple imputed datasets to replace missing values; 3) run regression on each imputed dataset; 4) run regression on non-imputed data with listwise deletion; 5) summarize imputed analysis and compare to non-imputed results.

RESULTS

1) Analyze missing value patterns: Utilizing PROC MI on variables {MPG, CYLINDERS, SIZE, HP, WEIGHT, ACCEL, ENG_TYPE}, we may visualize missing data patterns on our 38 records. As may be seen in Table 2, missing data appears unable to be re-ordered such that when a missing value is observed, all other values after that are also missing. Because of this, we determine that our data is arbitrary, or non-monotone. Interestingly, only 47.37% of the data is considered “complete” and without missing data.

Table 2 – Missing Data Patterns

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

With our dataset classified as non-monotone, we determine that the recommended imputation method is Markov Chain Monte Carlo (MCMC), the default method for PROC MI.

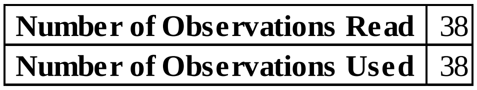
2) Create Multiple Imputed Datasets to Replace Missing Values: With the MCMC imputation method, we run PROC MI once more, with 5 imputations. With a seed value of ‘123’, utilized for reproducibility, we are able to compute a “miout” output dataset constituting the 5 imputations. As can be seen in Table 3, the multiple imputations executed successfully with inputs as described.

Table 3 – PROC MI 5 seeded imputations

Model Information	
Data Set	WORK.CARS
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	123

3) Run Regression on Each Imputed Data: In the previous step, PROC MI was used to create 5 imputed datasets (“miout”). In this step, PROC REG will be run on the “miout” dataset, meaning all 38 observations (shown in Table 4) will be included in each of the 5 datasets.

Table 4 – Imputed Regression, Read/Used Observations



Number of Observations Read	38
Number of Observations Used	38

Table 5 shows the parameter estimates per variable by imputation. This is a summary of each imputation to allow quick comparison of the parameter estimates across the imputed datasets.

Table 5 – Parameter Estimates for Multiple Imputation Regression by Imputation

Variable	1	2	3	4	5
Intercept	60.5684	58.9325	58.8236	59.3792	61.8267
CYLINDERS	-1.42743	-0.7811	-1.32468	-1.41455	-1.31251
SIZE	0.071662	0.056361	0.065822	0.065348	0.081335
HP	-0.03993	-0.07208	-0.03272	-0.05112	-0.07162
WEIGHT	-12.8238	-11.3978	-12.344	-11.7165	-12.9495

The output ("outreg") dataset includes parameter estimates and summaries of the regression results for the 5 imputed datasets. Each dataset was run through regression as a result of the BY statement included in the SAS code. This "outreg" data will be used by PROC MIANALYZE in the analysis step to compare regression with imputed values against the “original” regression using listwise deletion results.

4) Run Regression on Non-Imputed Data with Listwise Deletion: Applying listwise deletion on the dataset prior to running the regression function would result in only 22 of the total 38 observations being present in the final dataset, as seen in Table 6. This means 16 observations, representing 42% of the entire dataset, would be excluded from the analysis. Clearly this is an issue as the number of observations in the full dataset was limited to begin with.

Table 6 – Non-Imputed Regression, Read/Used Observations

Number of Observations Read	**38**
Number of Observations Used	22
Number of Observations with Missing Values	16

Table 7 shows the parameter estimates of regression using listwise deletion. These values serve as the "original" values for comparison with the imputed results.

Table 7 – Parameter Estimates for Non-Imputed Regression

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	59.29187	4.60156	12.89	<.0001
CYLINDERS	1	-1.52024	1.06901	-1.42	0.1731
SIZE	1	0.06595	0.02756	2.39	0.0285
HP	1	-0.06502	0.05948	-1.09	0.2895
WEIGHT	1	-10.66719	3.0213	-3.53	0.0026

From Table 7, both the size and weight variables are below the alpha threshold of 0.05, meaning they have statistical significance when predicting MPG.

5) Summarize Imputed Analysis and Compare to Non-Imputed Results: After producing estimates for each iteration of the imputed dataset and also regressing on the original listwise deletion dataset, the final step in the imputation process is to combine the results of each iteration using the PROC MIANALYZE statement and compare against listwise deletion results. Doing so produces the results indicated within the "Imputations" columns of Table 8. Provided are also the estimate results and statistics for the original dataset with listwise deletion as shown previously, now with the addition of 95% confidence interval values. Again, the original estimates were derived via regression performed on only 22 observations whereas the MCMC multiple imputation was performed on all 38 records.

Table 8 – Non-Imputed vs. Imputed Regression Results Comparison

	Parameter Estimates		Standard Errors		95% Confidence Intervals				Pr > |t|	
Variable	**Original**	**Imputations**	**Original**	**Imputations**	**Original**		**Imputations**		**Original**	**Imputations**
Intercept	59.2919	59.9061	4.6016	3.7391	49.5834	69.0003	52.5338	67.2784	<.0001	<.0001
CYLINDERS	-1.5202	-1.2521	1.0690	0.8406	-3.7757	0.7352	-2.9071	0.4030	0.1731	0.1375
SIZE	0.0660	0.0681	0.0276	0.0219	0.0078	0.1241	0.0247	0.1115	0.0285	0.0025
HP	-0.0650	-0.0535	0.0595	0.0427	-0.1905	0.0605	-0.1384	0.0314	0.2895	0.2137
WEIGHT	-10.6672	-12.2463	3.0213	2.5041	-17.0416	-4.2928	-17.1658	-7.3268	0.0026	<.0001

As Table 8 depicts, parameter estimates are very similar between MCMC imputation and listwise deletion methods; however, standard errors are slightly smaller across all imputed dataset attributes, and confidence intervals are in turn narrower. Given the parameter estimate similarities but narrower confidence intervals, imputed dataset p-values reflect increased statistical significance. While the difference in p-values is still minor and no differences cross the alpha threshold of 0.05, it is clear such a difference could violate this alpha value and, therefore, change one's interpretation for variable significance in MPG prediction.

DISCUSSION & FUTURE WORK

The original vehicle MPG dataset included 16 observations with missing data. When the common practice of listwise deletion is performed, we are affectively throwing away 42% of our data. Doing so introduces unwanted bias into our regression analysis. To mitigate the effects of such bias, multiple imputation algorithms such as those made available via SAS's PROC MI, CORR, and MIANALYZE may be utilized. Using these procedures to conduct MCMC multiple imputation methods enabled the use of all records rather than a subset based on missing data points. As such, this case study was an effective demonstration of multiple imputation principles as linear regression modeling for MPG prediction produced parameter estimates of greater statistical significance with smaller standard errors and narrower confidence intervals when imputation was performed.

As extension of this effort, future works include three items of interest. The first is to conduct the Fully Conditional Specification (FCS) regression or logistic regression methods to impute the missing data instead of MCMC since the CYLINDERS variable is ordinal and the ENG_TYPE variable is binary rather than continuous. FCS was not used for this case study as it was outside the scope of the lecture material and resources provided. The second item of interest for future work would be to clean up the final regression model by fine tuning the independent variable lineup used in regression modeling. This would be done by assessing Variable Inflation Factor (VIF) values, adjusted-R^2 values, and parameter estimate p-values with the inclusion or exclusion of the available independent variables. Finally, the third item of interest would be to replicate this case study using the R statistical programming language. Doing so would further validate our approach to imputation while providing different insights and algorithm options not available within SAS.

APPENDIX: SAS CODE

```
* Import csv data file ;
data CARS;
*infile '/folders/myshortcuts/SAS/carmpgdata_2.csv' dlm=',' DSD firstobs=2;
infile '/home/cboomhower0/sasuser.v94/MSDS7333/carmpgdata_2.csv' dlm=',' DSD
firstobs=2;
* SAS does not properly recognize empty values for delimited data unless you use the
dsd option. Need to use the dsd option on the infile statement if two consecutive
delimiters are used to indicate missing values (e.g., two consecutive commas, two
consecutive tabs). ;
input Auto $ MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE ;
run;
* Print imported carmpg data ;
proc print data=CARS;
run;

* Render scatterplot matrix and correlation matrix;
* aka Descriptive stats (EDA) for non-imputed data ;
title "Descriptive Stats /EDA for Non-imputed Data";
proc corr data=CARS;
* Important Note: PROC CORR will not perform listwise unless NOMISS specified ;
var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
run;

* Scatterplot matrix ;
proc sgscatter data=CARS;
matrix MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE / diagonal=(histogram normal);
run;

* Analyze missing patterns (monotone or non-monotone?) ;
title "Missing Data Pattern";
ods select misspattern;
proc mi data=CARS nimpute=0;
var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
run;

* Descriptive stats for listwise complete data ;
title "EDA on Listwise Complete Data";
proc corr data=CARS nomiss;
var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
run;

* Create imputed data files (specify nimpute=5) ;
title "Multiple Imputation (MI) Data";
ods graphics on;
proc mi data=CARS nimpute=5
out=miout seed=123;
var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
run;
* Print Multiple Imputation (MI) dataset output ;
```

```
title "MI out dataset";
proc print data=miout; run;

* Analyze Multiple Imputation (MI) Data ;
title "Linear Regression on Multiple Imputation (MI) Data";
proc reg data=miout outest=outreg covout;
model mpg = CYLINDERS SIZE HP WEIGHT;
by _Imputation_;
run;
* Print Multiple Imputation (MI) dataset output ;
title "Regression Output Data";
proc print data=outreg; run;

* Multiple Imputation Results Analysis ;
title "Multiple Imputation (MI) Results Analysis";
proc mianalyze data=outreg;
modeleffects CYLINDERS SIZE HP WEIGHT Intercept;
run;

* Analyze complete listwise data ;
title "Predicting MPG on Non Imputed Data (data with missing values) - Listwise
Deletion";
proc reg data=CARS;
 model mpg = CYLINDERS SIZE HP WEIGHT / CLB;   /* Display 95% CI */;
run;
```