Unit 2 Case Study – Multiple Imputation

**Zach Brown**

*MSDS 7333 – Quantifying the World, 5/24/2017*

**ABSTRACT**

The purpose of this case study is to analyze the impact of multiple imputation on the effectiveness of linear regression when run on an incomplete dataset. A regression model was first fit to the data using listwise deletion. Multiple imputation was then performed and a linear regression model was fit to the imputed dataset. The parameter estimates and standard errors for both datasets were then compared to one another. The regression model that was created from the imputed data produced must better results than the model created from the listwise deletion dataset.

***Keywords:*** Multiple Imputation, MCMC, Missing Data, Linear Regression, Fuel Economy, 38 Car Models

**Introduction**

When analyzing data, it is very common to encounter a dataset with missing values. Incomplete survey results, sensor failures, and human error are just three examples of the myriad reasons why a dataset might be incomplete. When analyzing a dataset, missing values can decrease the power of the analysis. Some methods cannot be used at all with incomplete data.

In this case study, we will examine a dataset containing data related to the fuel economy of cars. The features contained in this dataset include the make and model of the car, miles per gallon, cylinders, size, horsepower, weight, acceleration, and engine type. The data in this dataset is incomplete. The objective of this case study is compare the results of linear regression on this dataset before and after imputing the missing values and determine if multiple imputation had a meaningful impact on the analysis.

**BACKGROUND**

The dataset being used in this case study contains fuel economy data for 38 cars as measured in 2005. The variables included in this dataset are detailed in Table 1. This data also contains missing values. This will allow us to explore the impact of multiple imputation. To perform the analysis in this case study, we will use SAS version 9.4 TS Level 1M3.

|  |  |
| --- | --- |
| **Descriptions of Variables in Dataset** | |
| **Variable** | **Description** |
| Auto | Make and model of car |
| MPG | Estimated miles per gallon |
| CYLINDERS | Number of cylinders in engine |
| SIZE | Engine displacement |
| HP | Horsepower |
| WEIGHT | Weight of car |
| ACCEL | Acceleration |
| ENG\_TYPE | Engine type |

Table 1. Names and descriptions of the variables contained in the dataset

Multiple imputation is the main analysis method that will be examined in this case study. Multiple imputation “replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute” (Yuan). 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.

**METHODS**

The first step in performing this analysis was to perform linear regression on the original dataset. The dependent variable is MPG (miles per gallon or the fuel economy of the vehicle). The independent variables used are CYLINDERS, SIZE, HP (horsepower), WEIGHT, ACCEL (acceleration), and ENG\_TYPE (engine type).

Linear regression cannot be performed on an incomplete dataset. Therefore, to run the baseline linear regression analysis on this data, listwise deletion must carried out. This will delete every row with missing values. Using the PROC REG in SAS performs the listwise deletion automatically.

Before determining a method of imputation, we must first investigate the missing values and determine if there is a pattern. This is done using PROC MI in SAS. As can be seen in Table 2, the missing data is non-monotone (i.e. there is no discernible pattern of missingness).



Table 2. Missing data pattern analysis

The next step is to impute the missing data. Since the pattern of missingness is arbitrary, we can use the Markov Chain Monte Carlo (MCMC) method of multiple imputation. The SAS PROC MI using the MCMC method is used to create 5 imputed datasets. A seed of 35399 is used to enable the analysis to be reproduced.

Linear regression is now performed on each of the 5 imputed datasets. The same dependent and independent variables that were used in the initial linear regression using listwise deletion are again used when running linear regression on the imputed datasets.

The final step is to use SAS PROC MIANALYZE to summarize the results of the regression on the imputed datasets. We can then compare these results with the listwise deletion results and determine if multiple imputation had a noticeable effect on the outcomes.

**RESULTS**

The original dataset contains 38 observations. When using listwise deletion to prepare the data for the initial linear regression analysis, the number of observations is reduced to 18. More than half of the dataset contains missing values and is removed using listwise deletion. When employing multiple imputation instead of listwise deletion, we are able to include all 38 observations in the linear regression analysis instead of only 18.

The multiple imputation process using the Markov Chain Monte Carlo method produces the estimated mean values for the parameters detailed in Table 3. A 95% confidence interval is also provided for each mean estimate.



Table 3. Mean estimates for parameters using multiple imputation

After running the multiple imputation, it was then possible to perform linear regression on each of the 5 imputed datasets. Once that was done, we used SAS PROC MIANALYZE to produce combined estimates for the regression model parameters. These estimates can be found in Table 4 along with the parameter estimates produced by the linear regression analysis that was run on the listwise deletion version of the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | |
| **Variable** | **Original Estimate** | **Original Standard Error** | **Original p-value** | **Combined Estimate** | **Combined Standard Error** | **Combined p-value** |
| **Intercept** | 70.14772 | 8.03838 | <.0001 | 69.542738 | 4.676262 | <.0001 |
| **CYLINDERS** | -3.33403 | 1.56072 | 0.056 | -2.892369 | 0.766712 | 0.0002 |
| **SIZE** | 0.0228 | 0.03207 | 0.4918 | 0.030931 | 0.021663 | 0.1597 |
| **HP** | -0.19546 | 0.08065 | 0.0338 | -0.158924 | 0.046085 | 0.0013 |
| **WEIGHT** | -0.30623 | 5.13263 | 0.9535 | -3.214728 | 3.740139 | 0.4001 |
| **ACCEL** | -0.78199 | 0.58264 | 0.2066 | -0.721966 | 0.409842 | 0.103 |
| **ENG\_TYPE** | 6.5988 | 3.59008 | 0.0932 | 5.855301 | 1.579569 | 0.0002 |

Table 4. Comparison of regression model parameter estimates using listwise deletion and multiple imputation methods

It is fairly clear that the imputed data produces a better regression model than the listwise deletion data. For every variable including the intercept, the standard error and p-value are much smaller for the combined data than for the original data. This is logical, given that this dataset was relatively small to begin with and that the imputed datasets had more than twice as many observations than the listwise deletion dataset. This resulted in higher power for the analysis on the imputed data.

**CONCLUSION**

It is very rare for a dataset to be complete in the real world with no missing values. In many cases, this issue must be dealt with before it is worthwhile or even possible to perform the desired analysis on it. If the chosen analysis method can be performed on a dataset with missing values, it will most likely have lower power than if it were to be run on a complete dataset.

The objective of this study was to determine whether multiple imputation would produce better results than listwise deletion when performing linear regression on the given fuel economy dataset. This was indeed the case, as linear regression produced much better variable estimates when the imputed dataset was used than when listwise deletion was used.

**REFERENCES**

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). SAS Institute Inc, Rockville, MD, 49, 1-11.

**APPENDIX – SAS CODE**

/\* Zach Brown

MSDS 7333 - Quantifying the World

Unit 2 Case Study

5/24/2017 \*/

/\* Import data set with missing values \*/

**data** carmpg;

infile 'C:\Users\xzach\Google Drive\SMU\MSDS 7333 - Quantifying the World\Week 3\Unit 2 Case Study\carmpg.csv' dsd firstobs = **2**;

input Auto $ MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

/\* Perform linear regression on data set using listwise deletion \*/

**proc** **reg** data = carmpg;

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

**run**;

/\* Investigate missing data and determine if there is a pattern \*/

ods select misspattern;

**proc** **mi** data=carmpg nimpute=**0**;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

/\* Create 5 imputed data sets \*/

**proc** **mi** data=carmpg nimpute=**5**

out=miout seed=**35399**;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

/\* Perform linear regression on each of the 5 imputed data sets \*/

**proc** **reg** data=miout outest=outreg covout;

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

by \_imputation\_;

**run**;

/\* Combine the 5 regression models using proc mianalyze \*/

**proc** **mianalyze** data=outreg;

modeleffects CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE INTERCEPT;

**run**;