QTW Unit 2 Case Study

Multiple Imputation

Joshua Herrera, Matthew Payne, Remy Lagrois

2018

Abstract:

As a data scientist, the dream is to receive a data set free of errors; no strange characters, every entry being typed correctly, no misspellings, and no missing data. That never happens. Somebody cleans up the data, and it’s usually the data scientist. In this case study, we will be addressing the last point of the dream: missing data. Whether due to instrument error, user error, non-response, missing data is likely to be the most common type of issue one has with a data set. In our case study, we attempted to remedy the problem of missing data by using multiple imputation, a method in which missing values are replaced with values derived from others in its category. This method was applied to a data set of various cars, their miles per gallon they achieve, and some variables which are though to contribute to the car’s millage per gallon. We concluded that the regression conducted on the multiple imputed data sets was a more accurate estimate, based on the increased power of having more observations, than the regression preformed without multiple imputation (list-wise deletion).

Introduction:

The objective of this case study is to use multiple imputation on the Car MPG data set, which contains both continuous and discrete numerical data. The variables to be analyzed include Miles per Gallon (MPG), Cylinders, Size, Horse Power (HP), Weight, Accel, and Eng\_type. The data set given has multiple instances of missing data, and we will determine which options to use for multiple imputation, and compare the results with a list-wise deletion.

Literature Review:

NEED LITERATURE REVIEW

Methods:

The first step taken in this study was coming up with our baseline to which we will compare our final data set’s regression. This baseline was constructed by running a linear regression upon the initial data set, with MPG as our dependent variable. SAS preformed list-wise deletion, which discarded all rows in which there is missing data. After determining our baseline, we used the MI procedure to discover a non-monotone missing data pattern as shown in Figure 1. Because the data displays a non-monotone pattern of missing data, we used the default method for the MI procedure and output the results to a data set called MIcars.

The MI procedure created 5 different imputations as we directed it to with the nimpute= option. These 5 imputations are conglomerated into the output MIcars, and contains 190 records, 38 times our five imputations. Now that there were no missing values, we performed another linear regression, this time, with the MIcars output. The output of this regression went to another data set called carReg, and the regression was run five times, once for each imputation, with the by statement in SAS. Finally, the carReg output was analyzed with the MIanalyze procedure in SAS, to combine these results for a better and more balanced view of the data.

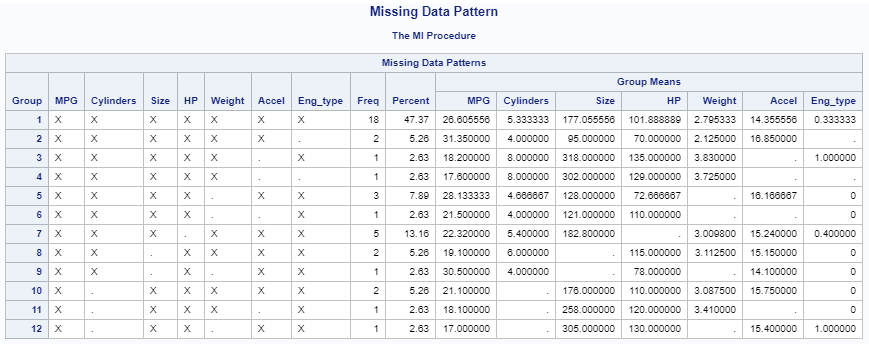


Figure 1

Results:

The preliminary list-wise deletion regression on the data set resulted in the parameter estimates shown in Figure 2. The model was determined to be:

MPG = 70.15 – 3.33 Cylinders + .02 Size - .20 HP - .31 Weight - .78 Accel + 6.60 Eng\_type

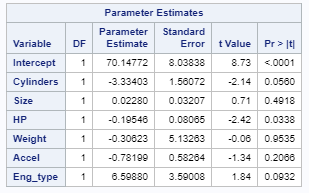


Figure 2

The most influential variables for this model are Eng\_type, which is weighted positively, and Cylinders, which is weighted negatively. Cars with more cylinders seem to have less MPG, and cars with engine type 1 appear to have more MPG. Acceleration has a negative correlation with MPG, with the rest of the variables having rather small parameter estimates less than half a point.

After imputing the data five times, regression was run on each new data set and the results can be found as Figure 3 below.

|  |  |
| --- | --- |
| Imputation 1 | Imputation 2 |
| Imputation 3 | Imputation 4 |
| Imputation 5 | |

Figure 3

While each of the five imputations are not perfect on their own, together, they should give us a better estimate of the population. Comparing these five imputations to the list-wise deletion regression, we see that the intercept is fairly consistently around 70. The Eng\_type variable has a strong positive weighting, and the Cylinders variable is roughly half of that in negative weight. It is notable that the Weight variable is a low positive number for some imputations like 1 and 3, but for the other imputations, it holds a strong negative weight. This can likely be attributed to the original list-wise deletion’s Weight variable having a massive standard error of 5.

The final step to this analysis is combining the five imputations. Figure 4 contains the parameter estimates for the combined regression. The final model is:  
MPG = 70.40 – 2.91 Cylinders + .03 Size - .17 HP – 2.75 Weight - .80 Accel + 5.25 Eng\_type

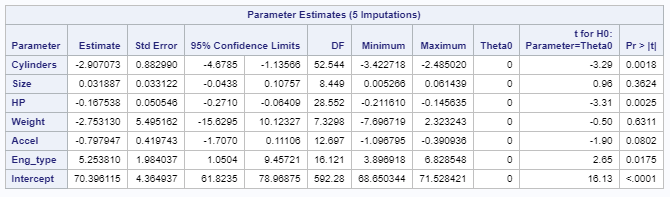


Figure 4

The final model is similar to the initial list-wise deletion regression in many ways; the Intercept, Size, Cylinders, HP, and Accel variables all have pretty similar estimates. Where the two models diverge are the Eng\_type, and Weight. Eng\_type decreased in its positive weight by approximately 1.35 score, and Weight decreased its score from -.31 to -2.75.

Conclusion:

NEED CONCLUSION  
Discuss the power? Need to add more pictures? ANOVA table?  
Discuss Std Error of Weight.

P-Values?

Appendix:

data cars;

infile '/home/herreraj0/Personal Folder/MSDS/MSDS7333 Quantifying the World/Case Study Week 2/carmpgdata\_2\_2\_2.txt' dsd dlm='09'x firstobs=2;

input Auto $ MPG Cylinders Size HP Weight Accel Eng\_type;

proc print data=cars;

run;

proc means data=cars NMISS N;

run;

title 'Regression with Listwise Deletion';

proc reg data=cars;

model mpg = Cylinders Size HP Weight Accel Eng\_type;

run;

title 'Missing Data Pattern';

ods select misspattern;

proc mi data=cars nimpute=0;

var MPG Cylinders Size HP Weight Accel Eng\_type;

run;

/\* seems non-monotone \*/

title 'Creating 5 Imputations';

proc mi data=cars out= MIcars nimpute=5 seed=3935;

var MPG Cylinders Size HP Weight Accel Eng\_type;

run;

title 'Conglomerated Imputated Data Sets';

proc print data=MIcars;

run;

title 'Regression Using Imputations';

proc reg data= MIcars outest= carReg covout;

model MPG = Cylinders Size HP Weight Accel Eng\_type;

by \_Imputation\_;

run;

title 'Combinging the Analyses';

proc mianalyze data=carReg;

modeleffects Cylinders Size HP Weight Accel Eng\_type Intercept;

run;