Authors: Justin Beer, Anthony Dalton, Harsha Kommanapalli

Date: 1/20/2018

**Analyzing Missing Data and**

**Multiple Imputation Models for Car MPG**

**Abstract**

The purpose of this case study is to compare the results obtained from a data set that has missing data utilizing a method of imputation versus a method of simple analysis on the missing data. Utilizing the number of cylinders, size, horsepower, and weight collected on various vehicles we will employ a linear regression model in order to determine the miles per gallon (mpg) for the vehicles. We will find that there are significant differences in the mpg model when using imputed versus missing data.

**Introduction**

The dataset that we are utilizing contains the vehicles make and model, miles per gallon, number of cylinders, engine size, horsepower, weight, acceleration, and engine type. The data is missing information on a missing completely at random basis. We determined the missing data is MCAR since there is no correlation between which data points are missing. Also, the pattern of missing-ness is arbitrary due to the data missing in no particular order. With this pattern of missing-ness in our data we will be using the method of MCMC full-data imputation in order to replace the missing values with estimates of the individual values.

**Literature Review**

The prevention and handling of the missing data by Hyun Kang [1] details practices and procedures to handle the inevitable problem that exists in research regarding missing data. Although the paper focuses on anesthesiology, all data scientists should be prepared and ready for missing-ness in some form. The best techniques to handle missing data are to properly plan the study and data collection properly. This can be achieved by limiting the number of study participants, using proper screening methods, personnel training, proper piloting, level-setting, participant engagement, and proper handling of survey participants withdraw. With big and more readily available data, often the person doing the data analysis doesn’t have the opportunity to lead the data collection process. Practices like listwise, casewise, and pairwise deletion are outlined as strategies but are not optimal practices as many records and results would be eliminated. The accepted practice with large enough sets of data is multiple imputation. Multiple imputation replaces missing values replaced by sets of plausible values from other variables in the dataset. With modern software packages and computing power, big datasets can be handled properly, and multiple imputation can be utilized quickly and easily. This is the approach that is utilized in our current study for Car MPG with missing data.

**Methods**

The first method we perform is a linear regression analysis of the dataset [We read in provided data into WORK.CarMpgclass] that has the missing values. In order to analyze this data set we used the following code:

PROC REG DATA = WORK.CarMpgclass;

MODEL MPG = CYLINDERS SIZE HP WEIGHT;

RUN;

The model that we are analyzing is the miles per gallon (MPG) of the vehicle as a function of the vehicles number of cylinders (CYLINDERS), engine size (SIZE), engine horsepower (HP), and vehicle weight (WEIGHT). From this analysis, which includes the missing data, we obtain Tables 1-4.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 927.64081 | 231.91020 | 37.77 | <.0001 |
| **Error** | 17 | 104.37374 | 6.13963 |  |  |
| **Corrected Total** | 21 | 1032.01455 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2.47783 | **R-Square** | 0.8989 |
| **Dependent Mean** | 26.24545 | **Adj R-Sq** | 0.8751 |
| **Coeff Var** | 9.44098 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **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 | 3.02130 | -3.53 | 0.0026 |

Tables 1-4: PROC REG output from dataset containing missing values.

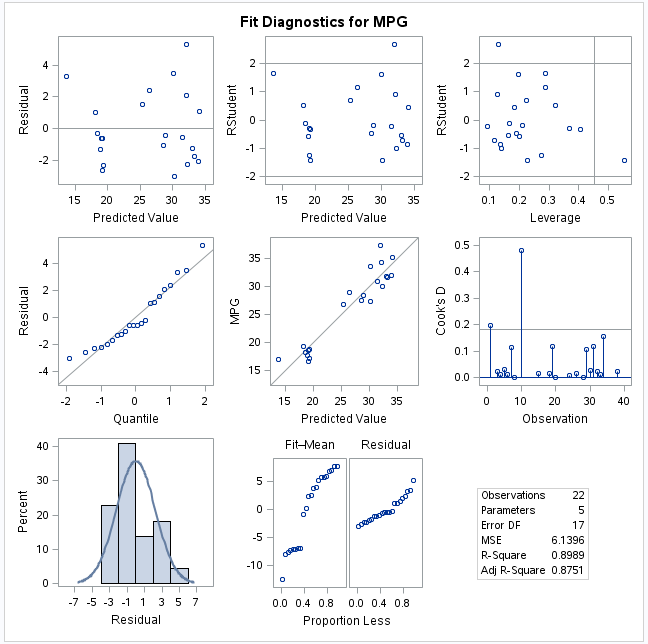


Figure 2: Fit diagnostics for Linear Regression model.

As we can see from Tables 1-4, there were 16 observations that contained missing values. SAS by default utilizes list-wise deletion which means it completely omits those 16 rows of observations from the linear regression analysis. The total number of observations used to create this linear regression model is only 22.

Even with this small amount of observations the model produced has an adjusted R2 = 0.8751. However, we must be careful to accept this model as list-wise deletion creates bias in the results and the model, while fitting the non-missing data quite well may not fit further data well. We can also see that the meaningful parameter estimates for this model are the SIZE and WEIGHT variables. Each has a p-value < 0.05.

In order to discover the pattern of missing-ness in our data we utilize the following code:

ODS select misspattern;

proc mi data = WORK.CarMpgclass nimpute= 0;

var CYLINDERS SIZE HP WEIGHT MPG;

RUN;

From this code we obtain Table 5 below. We see that the pattern of missing-ness is non- monotone, arbitrary with values missing completely at random.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Missing Data Patterns** | | | | | | | | | | | | |
| **Group** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **MPG** | **Freq** | **Percent** | **Group Means** | | | | |
| **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **MPG** |
| **1** | X | X | X | X | X | 22 | 57.89 | 5.454545 | 181.681818 | 101.727273 | 2.823682 | 26.245455 |
| **2** | X | X | X | . | X | 4 | 10.53 | 4.500000 | 126.250000 | 82.000000 | . | 26.475000 |
| **3** | X | X | . | X | X | 5 | 13.16 | 5.400000 | 182.800000 | . | 3.009800 | 22.320000 |
| **4** | X | . | X | X | X | 2 | 5.26 | 6.000000 | . | 115.000000 | 3.112500 | 19.100000 |
| **5** | X | . | X | . | X | 1 | 2.63 | 4.000000 | . | 78.000000 | . | 30.500000 |
| **6** | . | X | X | X | X | 3 | 7.89 | . | 203.333333 | 113.333333 | 3.195000 | 20.100000 |
| **7** | . | X | X | . | X | 1 | 2.63 | . | 305.000000 | 130.000000 | . | 17.000000 |

Table 5: Missing Data Patterns.

Now that we have established the pattern of missing-ness and analyzed our dataset initially, we can now move onto imputing our missing data values, creating a new dataset from the imputations, and analyzing the new dataset so that we may compare the effects between list-wise deletion and multiple imputation.

In order to impute our dataset we utilize the following code:

PROC MI data=WORK.CarMpgclass out=MICarMpgOut seed=33333;

var CYLINDERS SIZE HP WEIGHT MPG;

RUN;

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variance Information (25 Imputations)** | | | | | | | |
| **Variable** | **Variance** | | | **DF** | **Relative Increase**  **in Variance** | **Fraction Missing Information** | **Relative Efficiency** |
| **Between** | **Within** | **Total** |
| **CYLINDERS** | 0.000960 | 0.067572 | 0.068571 | 34.627 | 0.014782 | 0.014584 | 0.999417 |
| **SIZE** | 0.736779 | 208.503251 | 209.269501 | 35.021 | 0.003675 | 0.003663 | 0.999854 |
| **HP** | 0.373029 | 18.853302 | 19.241252 | 34.421 | 0.020577 | 0.020196 | 0.999193 |
| **WEIGHT** | 0.000109 | 0.013139 | 0.013253 | 34.846 | 0.008611 | 0.008544 | 0.999658 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates (25 Imputations)** | | | | | | | | | | |
| **Variable** | **Mean** | **Std Error** | **95% Confidence Limits** | | **DF** | **Minimum** | **Maximum** | **Mu0** | **t for H0: Mean=Mu0** | **Pr > |t|** |
| **CYLINDERS** | 5.429151 | 0.261860 | 4.8973 | 5.9610 | 34.627 | 5.368327 | 5.481464 | 0 | 20.73 | <.0001 |
| **SIZE** | 179.586802 | 14.466150 | 150.2196 | 208.9540 | 35.021 | 178.155935 | 181.178161 | 0 | 12.41 | <.0001 |
| **HP** | 102.327051 | 4.386485 | 93.4167 | 111.2374 | 34.421 | 101.142311 | 103.361877 | 0 | 23.33 | <.0001 |
| **WEIGHT** | 2.866895 | 0.115120 | 2.6332 | 3.1006 | 34.846 | 2.850679 | 2.888938 | 0 | 24.90 | <.0001 |

Tables 6-7: Variance Information and Parameter Estimates from PROC MI, imputed data.

From imputing our dataset we obtained parameter estimates that we used to replace the missing values. All of the parameter estimates, using 25 imputations, have a p-value < 0.0001. We can now analyze our imputed dataset using linear regression:

PROC REG data=MICarMpgOut outest=OutCarMpgReg covout;

MODEL MPG = CYLINDERS SIZE HP WEIGHT;

by \_imputation\_;

RUN;

Now that we have ran our linear regression model on all 25 of our imputed datasets we can now combine the results with our initial dataset results containing missing values:

PROC MIANALYZE data=OutCarMpgReg;

MODELEFFECTS CYLINDERS SIZE HP WEIGHT Intercept;

RUN;

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates (25 Imputations)** | | | | | | | | | | |
| **Parameter** | **Estimate** | **Std Error** | **95% Confidence Limits** | | **DF** | **Minimum** | **Maximum** | **Theta0** | **t for H0: Parameter=Theta0** | **Pr > |t|** |
| **CYLINDERS** | -1.573494 | 0.845389 | -3.2326 | 0.08561 | 923.09 | -2.055412 | -0.614378 | 0 | -1.86 | 0.0630 |
| **SIZE** | 0.075133 | 0.020995 | 0.0339 | 0.11632 | 1147.1 | 0.043939 | 0.083858 | 0 | 3.58 | 0.0004 |
| **HP** | -0.052521 | 0.040968 | -0.1330 | 0.02796 | 526.4 | -0.089257 | -0.005882 | 0 | -1.28 | 0.2004 |
| **WEIGHT** | -12.36336 | 2.430524 | -17.1335 | -7.59324 | 904.21 | -13.883779 | -9.846040 | 0 | -5.09 | <.0001 |
| **Intercept** | 60.642383 | 3.710887 | 53.3605 | 67.92427 | 1015.5 | 55.399271 | 62.263521 | 0 | 16.34 | <.0001 |

Table 8: Parameter Estimates from PROC MIANALYZE, imputed data.

**Results**

The initial dataset with missing values obtained the following biases linear regression model. The model had an adjusted R2 = 0.8751 using 22 observations. Only the SIZE and WEIGHT variables were statistically relevant with p-values < 0.05.

The following is the model produced by using 25 imputations on our dataset. We find that the SIZE and WEIGHT are again the only statistically relevant variables.

The results are what we expected. We did not expect that the statistical significance of the variables would materially change. This is due to the domain understanding that the SIZE and WEIGHT are the most significant variables. The CYLINDERS and HP are highly correlated with the engines SIZE.

**Conclusion**

While our missing value and multiple imputations models are very similar we can rest assured that using the multiple imputations method has resulted in a model with no bias caused by missing values, that accurately reflects missing value uncertainty, and that utilizes a representative random sample of the missing values.

In order to move forward with creating a better model for the vehicles mpg we need to address the issue of why was there missing values in our dataset. When possible, it is always best to fully collect the information we want to analyze. While this is not always feasible, we should attempt to adjust our data collection process in order to control for missing values as best we can. For our dataset we could contact manufacturers directly as they will always have their vehicles make/model, engine size, engine type, horsepower rating, and weight.

**Appendix**

|  |  |
| --- | --- |
| [1] | H. Kang, The prevention and handling of the missing data, Korean Journal of Anesthesiology, 2013. |