Case Study 2 - Multiple Imputation: Car MPG

Yao Yao, Brian Kruse, Ivelin Angelov

MSDS 7333 Section 402

May 21, 2018

**1 Introduction**

Missing data from the cause of various errors in data collection or corruption in dataset can lead to unwanted biases in the analysis and modeling. Some software only analyze observations with non-missing data rows, which further reduces the amount of original data the software actually uses. In order to reduce bias and increase the study’s statistical power, multiple imputation via Markov Chain Monte Carlo (MCMC) was used to create five imputed datasets to create a better linear model for miles per gallon (MPG) based on correlated explanatory variables [1] [2].

The objective of this paper is to show the methods needed to obtain a better model with unbiased estimates and increased statistical power with imputation than the original dataset after listwise deletion.

**2 Methods**

The steps used for this analysis were data exploratory analysis using correlation, summary statistics, and matrix plots. Once the correlation coefficients, data types, and linear trends are identified, use missing value pattern diagnostics to see which type of imputation is needed to salvage more of the observations with missing values. Create multiple imputed datasets to replace missing values and run regression models on each of the imputed datasets. Aggregate the imputed regression analysis and compare those parameter statistics with that of the original non-imputed dataset after listwise deletion.

**3 Exploratory Data Analysis**

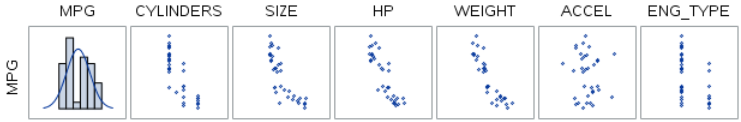
The Car MPG dataset contains 38 total observations with 8 variables, where the miles per gallon is the target variable and is dependent on the following explanatory variables: number of cylinders, engine size, horsepower, vehicle weight, vehicle acceleration, and engine type. Only the vehicle model and miles per gallon are complete with no missing variables while the rest of the explanatory variables have a total of 25 missing values. Table 1 shows the basic statistics summary as well as data descriptions and type, where the variables are classified as string, double, integer, binary integer, and even integer (because it is rare for common car models to have an odd number of cylinders). Vehicle model is not a numerical variable and will not be used for the imputation model but could be used later to look up the actual values of the missing numbers to validate the imputation process.

**Table 1**: Basic Statistics Summary of The Car MPG Dataset: Data Descriptions and Type

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **Data Type** | **Variable** | **NAs** | **Min** | **Q1** | **Mean** | **Med** | **Std Dev** | **Q3** | **Max** |
| Auto | Vehicle Model | String | Data Record | 0 |  |  |  |  |  |  |  |
| MPG | Miles Per Gallon | Double | Target | 0 | 15.5 | 18.5 | 24.76 | 24.25 | 6.54 | 30.5 | 37.3 |
| CYLINDERS | # of Cylinders | Even Integer | Explanatory | 4 | 4 | 4 | 5.32 | 4 | 1.6 | 6 | 8 |
| SIZE | Engine Size | Integer | Explanatory | 3 | 85 | 105 | 180.88 | 151 | 91.42 | 258 | 360 |
| HP | Horsepower | Integer | Explanatory | 5 | 65 | 75 | 101.33 | 97 | 27.11 | 125 | 155 |
| WEIGHT | Vehicle Weight | Double | Explanatory | 6 | 1.915 | 2.265 | 2.9 | 2.747 | 0.709 | 3.507 | 4.36 |
| ACCEL | Vehicle Acceleration | Double | Explanatory | 4 | 11.3 | 14.1 | 14.94 | 14.8 | 1.58 | 15.8 | 19.2 |
| ENG\_TYPE | Engine Type | Binary Integer | Explanatory | 3 | 0 | 0 | 0.285 | 0 | 0.45 | 1 | 1 |

The scatter plot matrix shows that the target variable, MPG, has a significant negative correlation Pearson's coefficient with the number of cylinders, engine size, horsepower, vehicle weight, and engine type, where those P values are less than 0.05 [Fig. 1]. Only the explanatory variable acceleration does not have a significant P value when evaluating correlation to MPG [Fig. 1]. Due to these linear trends, those explanatory variables that have significant correlation with MPG may result in more accurate numbers than that for vehicle acceleration in later modeling.

**Figure 1**: Scatter Matrix Plot of The Car MPG Dataset with Pearson's Correlation and P Value



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MPG** | **MPG** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **ACCEL** | **ENG\_TYPE** |
| Correlation | 1 | -0.82355 | -0.8163 | -0.87009 | -0.90284 | -0.10693 | -0.4583 |
| P Value |  | <.0001 | <.0001 | <.0001 | <.0001 | 0.5472 | 0.0056 |
| # of Obs | 38 | 34 | 35 | 33 | 32 | 34 | 35 |

**4 Results**

**1) Use PROC MI to discover the missing values patterns and to decide what MI options to use. (Assume no need for transformations.)**

**Table 2**: The Visualization of Missing Data Patterns for the Car MPG Dataset

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **MPG** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **ACCEL** | **ENG\_TYPE** | **Freq** | **Percent** | **NAs Per Group** |
| 1 | X | X | X | X | X | X | X | 18 | 47.37 | 0 |
| 2 | X | X | X | . | X | X | X | 5 | 13.16 | 1 |
| 3 | X | X | X | X | . | X | X | 3 | 7.89 | 1 |
| 4 | X | X | X | X | X | X | . | 2 | 5.26 | 1 |
| 5 | X | X | . | X | X | X | X | 2 | 5.26 | 1 |
| 6 | X | . | X | X | X | X | X | 2 | 5.26 | 1 |
| 7 | X | X | X | X | X | . | X | 1 | 2.63 | 1 |
| 8 | X | X | X | X | X | . | . | 1 | 2.63 | 2 |
| 9 | X | X | X | X | . | . | X | 1 | 2.63 | 2 |
| 10 | X | X | . | X | . | X | X | 1 | 2.63 | 2 |
| 11 | X | . | X | X | X | . | X | 1 | 2.63 | 2 |
| 12 | X | . | X | X | . | X | X | 1 | 2.63 | 2 |

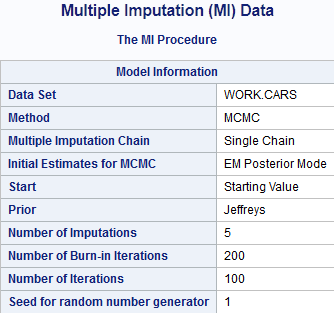
When the missing data patterns are visualized for the Car MPG dataset, it is unable to be re-ordered so that when a missing value is observed, all other values after that are also missing [Table 2]. Therefore, the dataset’s pattern of missing data was found to be arbitrary and non-monotone. The method of imputation based on the non-monotone patterns of missing data is Markov Chain Monte Carlo (MCMC) full-data imputation via PROC MI, using the default option on SAS [1] [2].

18 of the 38 observations (47.37%) have no missing values, 15 of the 38 observations (39.47%) have 1 missing value, and 5 of the 38 observations (13.17%) have 2 missing values [Table 2]. There are no observations with 3 or more missing values.

**2) Use PROC MI to create multiple imputed data sets.**

Using the default settings of Markov Chain Monte Carlo full-data imputation via PROC MI for 5 imputation cycles and a seed value of 1, 190 observations (38\*5) were created [Table 3].

**Table 3**: Summary Procedure of Markov Chain Monte Carlo via PROC MI for 5 Imputation Cycles



Upon looking at the last imputation cycle of the imputed table, we see that the imputed data has excess decimals for explanatory variables that should be integers including number of cylinders, size, horsepower, and engine type [Table 4]. They are not rounded to the nearest integer in order to keep the variance and parameter estimates consistent with the MCMC outputs for later regression analysis. We would round those numbers prior to checking them for accuracy if given the completed real dataset.

**Table 4:** Highlighted Are Imputed Variables for the Cars MPG Dataset After the 5th Imputation Cycle

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Auto** | **MPG** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **ACCEL** | **ENG\_TYPE** |
| Ford Country Sq. Wagon | 15.5 | 8 | 351 | 141.5 | 4.054 | 14.3 | 1 |
| Dodge Omni | 30.9 | 4 | 105 | 75 | 2.23 | 14.5 | -0.34 |
| Audi 5000 | 20.3 | 5 | 131 | 105.13 | 2.83 | 15.9 | 0 |
| Saab 99 GLE | 21.6 | 4.66 | 121 | 115 | 2.795 | 15.7 | 0 |
| Peugeot 694 SL | 16.2 | 6 | 191.38 | 133 | 3.41 | 15.8 | 0 |
| Buick Century Spec. | 20.6 | 6.18 | 231 | 105 | 3.38 | 15.8 | 0 |
| Mercury Zephyr | 20.8 | 6 | 200 | 85 | 2.76 | 16.7 | 0 |
| AMC Concord D/L | 18.1 | 6.94 | 258 | 120 | 3.41 | 14.64 | 0 |
| Chevy Caprice Classic | 17 | 8.36 | 305 | 130 | 3.76 | 15.4 | 1 |
| Ford LTD | 17.6 | 8 | 302 | 129 | 3.725 | 14.88 | 1.04 |
| Dodge St Regis | 18.2 | 8 | 318 | 135 | 3.83 | 13.62 | 1 |
| Ford Mustang 4 | 26.5 | 4 | 140 | 97.19 | 2.585 | 14.4 | 0 |
| Ford Mustang Ghia | 21.9 | 6 | 171 | 110.87 | 2.91 | 16.6 | 1 |
| Mazda GLC | 34.1 | 4 | 86 | 65 | 2.08 | 15.2 | 0 |
| AMC Spirit | 27.4 | 4 | 121 | 91.95 | 2.67 | 15 | 0 |
| Honda Accord | 29.5 | 4 | 98 | 68 | 2.24 | 16.6 | 0 |
| Datsun 210 | 31.8 | 4 | 85 | 65 | 2.02 | 19.2 | 0.29 |
| VW Dasher | 30.5 | 4 | 101.51 | 78 | 2.36 | 14.1 | 0 |
| Datsun 810 | 22 | 6 | 152.87 | 97 | 2.815 | 14.5 | 0 |
| BMW 320i | 21.5 | 4 | 121 | 110 | 2.65 | 17.26 | 0 |

**3) Use PROC REG to analyze the multiple data sets while outputting information to be used in MIANALYZE.**

Using the 5 sets of imputation generated from PROC MI's MCMC, 5 sets of regression analysis for explanatory variables were created, which includes number of cylinders, engine size, horsepower, vehicle weight, and engine type, where their P values for linear regression were all significant [Table 5]. The explanatory variable acceleration was not used in the regression model because earlier correlation coefficient testing proved it was not significant and that variable also did not contribute to the model in significance [Table 5, Fig. 1].

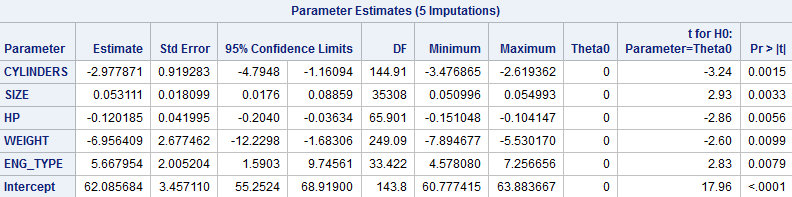
**Table 5**: 5 Sets of Regression Analysis for Significant Explanatory Variables Created from Imputation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Imputation Cycle** | **Variable** | **Param Est** | **Std Err** | **t Value** | **Pr > |t|** |
| 1 | CYLINDERS | -3.16493 | 0.75054 | -4.22 | 0.0002 |
| 2 | CYLINDERS | -3.47686 | 0.8999 | -3.86 | 0.0005 |
| 3 | CYLINDERS | -2.61936 | 0.88017 | -2.98 | 0.0055 |
| 4 | CYLINDERS | -2.78365 | 0.82965 | -3.36 | 0.0021 |
| 5 | CYLINDERS | -2.84454 | 0.829 | -3.43 | 0.0017 |
| 1 | SIZE | 0.05499 | 0.01582 | 3.48 | 0.0015 |
| 2 | SIZE | 0.051 | 0.01823 | 2.8 | 0.0086 |
| 3 | SIZE | 0.05168 | 0.01845 | 2.8 | 0.0086 |
| 4 | SIZE | 0.05425 | 0.01839 | 2.95 | 0.0059 |
| 5 | SIZE | 0.05363 | 0.01895 | 2.83 | 0.008 |
| 1 | HP | -0.11696 | 0.03127 | -3.74 | 0.0007 |
| 2 | HP | -0.15105 | 0.03854 | -3.92 | 0.0004 |
| 3 | HP | -0.12335 | 0.03601 | -3.43 | 0.0017 |
| 4 | HP | -0.10415 | 0.03451 | -3.02 | 0.005 |
| 5 | HP | -0.10542 | 0.04116 | -2.56 | 0.0154 |
| 1 | WEIGHT | -7.11299 | 2.13231 | -3.34 | 0.0022 |
| 2 | WEIGHT | -5.53017 | 2.4515 | -2.26 | 0.031 |
| 3 | WEIGHT | -6.99136 | 2.48855 | -2.81 | 0.0084 |
| 4 | WEIGHT | -7.89468 | 2.67911 | -2.95 | 0.0059 |
| 5 | WEIGHT | -7.25284 | 2.71565 | -2.67 | 0.0118 |
| 1 | ENG\_TYPE | 6.22165 | 1.55912 | 3.99 | 0.0004 |
| 2 | ENG\_TYPE | 7.25666 | 1.61757 | 4.49 | <.0001 |
| 3 | ENG\_TYPE | 5.32117 | 1.61326 | 3.3 | 0.0024 |
| 4 | ENG\_TYPE | 4.96222 | 1.73675 | 2.86 | 0.0075 |
| 5 | ENG\_TYPE | 4.57808 | 1.57571 | 2.91 | 0.0066 |
| 1 | Intercept | 62.72736 | 2.70442 | 23.19 | <.0001 |
| 2 | Intercept | 63.88367 | 3.61247 | 17.68 | <.0001 |
| 3 | Intercept | 60.77741 | 3.11338 | 19.52 | <.0001 |
| 4 | Intercept | 62.08934 | 3.10394 | 20 | <.0001 |
| 5 | Intercept | 60.95064 | 3.17806 | 19.18 | <.0001 |

**4) Use PROC MIANALYZE to summarize the imputed analyses.**

The output of the PROC REG data is fed into PROC MIANALYZE to summarize the regression analysis for those explanatory variables from Table 5 into Table 6, where all parameter estimates are aggregated and are significant.

**Table 6**: Summary of Regression Analysis for Significant Explanatory Variables Created from 5 Sets of Imputation



**5) Compare these results to the listwise deletion results.**

To create a linear regression model based on listwise deletion, only the significant variables from linear regression after imputation were used for the proper comparison. Therefore, all explanatory variables except acceleration were used to build the listwise deletion regression model. Table 7 shows that 19 of 38 (50%) of the observations do not have missing values because Group 7 from Table 2 shows that there is one observation that is only missing acceleration but that observation is considered complete due to the linear regression model not using that variable [Table 7].

**Table 7**: 19 Instead of 18 Observations are Considered 'Complete' Because the Model No Longer Uses Acceleration



Although the parameter estimates of the explanatory variables are similar for possibly creating the same model when predicting MPG, the P values of the original dataset using listwise deletion shows that 4 of the 5 explanatory variables are insignificant towards MPG whereas all 5 of 5 explanatory variables are significant for the imputed model [Table 8]. Therefore, imputation is important especially for a small dataset where half of the observations were used to create the model due to missing values.

**Table 8**: Comparison of Regression Estimates with Imputed Values Against the Original Regression Using Listwise Deletion

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Estimate** | | **Std Error** | | **95% Confidence Limits** | | | | **Pr > |t|** | |
| **Parameter** | **Imputed** | **Original** | **Imputed** | **Original** | **Imputed** | | **Original** | | **Imputed** | **Original** |
| CYLINDERS | -2.977871 | -3.46451 | 0.919283 | 1.5268 | -4.7948 | -1.16094 | -6.76296 | -0.16606 | 0.0015 | 0.0409 |
| SIZE | 0.053111 | 0.03464 | 0.018099 | 0.03031 | 0.0176 | 0.08859 | -0.03084 | 0.10013 | 0.0033 | 0.2737 |
| HP | -0.120185 | -0.13252 | 0.041995 | 0.0654 | -0.204 | -0.03634 | -0.2738 | 0.00877 | 0.0056 | 0.0638 |
| WEIGHT | -6.956409 | -4.05552 | 2.677462 | 4.24187 | -12.2298 | -1.68306 | -13.21952 | 5.10847 | 0.0099 | 0.3565 |
| ENG\_TYPE | 5.667954 | 7.42548 | 2.005204 | 3.51425 | 1.5903 | 9.74561 | -0.1666 | 15.01756 | 0.0079 | 0.0545 |
| Intercept | 62.085684 | 61.29123 | 3.45711 | 4.5608 | 55.2524 | 68.919 | 51.43821 | 71.14424 | <.0001 | <.0001 |

Furthermore, the standard errors are smaller for all imputed dataset attributes, which would create narrower confidence intervals [Table 8]. Having a P value that is not significant for the original model would possibility result in the removal of certain features to create a different final regression estimate and result.

**5 Conclusion**

The original Car MPG dataset had 20 of 38 (52.63%) observations with missing data, where SAS would only read in 18 (47.37%) of the observations, which introduces unwarranted biases when creating analysis and models. Correlation, summary statistics, and matrix plots allowed the necessary initial data exploration when proving which explanatory variables were significant for the linear regression model. Missing value pattern diagnostics was able to pinpoint the type of imputation needed in order to salvage the rest of the dataset with missing values. Once 5 cycles of imputed numbers are generated and summarized, the new linear regression model created better estimates because bias was reduced statistical power was increased from having the completed dataset [2].

Having an imputed dataset would also lower standard errors for smaller confidence intervals and change variable significance. Variable significance not only changes the interpretation of the model but also create a different regression model if those features were removed. Therefore, conducting MCMC multiple imputation enabled the use of all records, which resulted in a better regression model.

**References**  
  
1) D. B. Rubin, “Basic Ideas of Multiple Imputation for Nonresponse,” Survey Methodology, vol. 12, no.1, pp. 37-47, June 1986.

2) T. Rosenström, “Lecture Notes: Some Core Ideas of Imputation for Nonresponse in Surveys,” Univ. Helsinki, FI, May 14, 2014.

**Appendix: SAS Code**

FILENAME REFFILE '/home/yaoy890/carmpgdata\_2.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=CARS;

GETNAMES=YES;

RUN;

\*EDA;

proc means data=CARS min q1 mean median q3 max;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

run;

title "EDA for Non-imputed Data";

proc corr data=CARS;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE ;

run;

proc sgscatter data=CARS;

matrix MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE / diagonal=(histogram normal);

run;

title "Scatterplot";

\*Part 1;

proc mi data=CARS nimpute=1;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

run;

title "EDA on Listwise Complete Data";

proc corr data=CARS nomiss;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

run;

\*Part 2;

title "Multiple Imputation (MI) Data";

proc mi data=CARS nimpute=5 out=out seed=1;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

run;

title "MI out dataset";

proc print data=out;

run;

\*Part 3;

title "Linear Regression on Multiple Imputation (MI) Data";

proc reg data=out outest=outreg covout;

model MPG = CYLINDERS SIZE HP WEIGHT ENG\_TYPE;

by \_Imputation\_;

run;

title "Regression Output Data";

proc print data=outreg;

run;

\*Part 4;

title "Multiple Imputation (MI) Results Analysis";

proc mianalyze data=outreg;

modeleffects CYLINDERS SIZE HP WEIGHT ENG\_TYPE Intercept;

run;

\*Part 5;

title "Predicting MPG on Non Imputed Data - Listwise Deletion";

proc reg data=CARS;

model mpg = CYLINDERS SIZE HP WEIGHT ENG\_TYPE / CLB;

run;