Multiple Imputation for Calculating Car’s MPG

**Abstract**

This report contains a detailed guide to performing multiple imputation on a data set containing missing values. It contains background information about why data could be imputed and other methods of dealing with missing data. It will detail results using SAS functions such as PROC REG, PROC MI, and PROC MIANALIZE. PROC REG will be used to measure how improved the data is after imputing the data.

**Introduction**

We are looking at the miles per gallon observed from 38 different automobiles. The factors that were observed were the miles per gallon, the number of cylinders in the vehicle, the size of the engine, the horsepower, the weight of the automobile, acceleration, and engine type. For the size of the engine, it is calculated in such a way that the larger the number, the larger the engine.

One issue with our data is that of missing values in 19 of our 38 cases. Before we conduct our analysis, we will need to impute data into our data set. The reason for this is that as the data is now, SAS will not analyze cases that are missing any data. This means that if only 1 out of 7 of our pieces of data are missing from a case, SAS will not use any of the data from that case. When imputed though, we will be able to analyze our data. Therefore, we will impute data to gather a more accurate measure of the miles per gallon of each car.

**Literature Review**

There are 3 common types of missing data. missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data that is missing completely at random do not have any correlation between missing data and data that is not missing. Data that is missing at random is data that is either available or missing depending on a conditional statement regarding a different variable. Data that is missing not at random is data that is systematically or purposely withheld or not gathered for the data set.

Some of the common methods for dealing with missing data are complete case analysis, pairwise deletion, estimation of individual missing values, and estimation of missing value characteristics. Complete case analysis, also known as listwise deletion, uses only records that do not have any missing data. This means that if even one attribute of a record is missing out of 10, then none of the values, including the available 9 are to be used. The issue with complete case analysis is that this reduces the sample size by using less records than we have available to us. This could cause bias in the data which then could lead to analysis drawing inaccurate representations about what the data is. Complete case analysis is the standard method used by SAS when dealing with missing values unless otherwise programmed.

Pairwise deletion can only be used when any missing data is considered to be missing completely at random. Also known as available cases analysis, it, unlike complete case analysis, will use all the data available to it and not use complete records of data because of any missing data it may contain. Pairwise deletion is used in SAS when using the PROC CORR function.

Other methods for dealing with missing data are single imputation, multiple imputation, mean substitution, dummy variables, and regression substitution. Single imputation fills in each missing value in the data set with a method. Some of these methods are mean substitution, regression substitution, and dummy variables. Multiple imputation is the process of simple imputation done many times and then the results are averaged. In multiple imputation, a new data set is created for each different set of imputations.

**Methods Section**

Before we impute any missing data, we must first determine what type of missing data that we have (MCAR, MAR, or MNAR). Then, we will use PROC REG in order to have an initial analysis. Then, PROC MI will be used to create a set multitude of data sets that were each imputed. If the pattern of the missing data is considered arbitrary, we will use the MCMC method. After we have created the imputed data, we will analyze the imputed data using PROC REG. We can then check these results against our initial regression results from our first PROC REG and see what change has occurred.

**Results Section**

We will first perform a regression for our initial analysis. Our equation that we will be using is MPG + CYLINDERS + SIZE + HP + WEIGHT + ACCEL + ENG\_TYPE.

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 18 |
| **Number of Observations with Missing Values** | 20 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 774.27999 | 129.04667 | 22.39 | <.0001 |
| **Error** | 11 | 63.40945 | 5.76450 |  |  |
| **Corrected Total** | 17 | 837.68944 |  |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 70.14772 | 8.03838 | 8.73 | <.0001 |
| **CYLINDERS** | **1** | -3.33403 | 1.56072 | -2.14 | 0.0560 |
| **SIZE** | **1** | 0.02280 | 0.03207 | 0.71 | 0.4918 |
| **HP** | **1** | -0.19546 | 0.08065 | -2.42 | 0.0338 |
| **WEIGHT** | **1** | -0.30623 | 5.13263 | -0.06 | 0.9535 |
| **ACCEL** | **1** | -0.78199 | 0.58264 | -1.34 | 0.2066 |
| **ENG\_TYPE** | **1** | 6.59880 | 3.59008 | 1.84 | 0.0932 |

In our initial analysis, we can see that only 18 of the 38 records were used. This means that there are 20 records in the dataset that have at least 1 missing piece of data. This gives us a corrected total or power of 17. This analysis gives the following equation using the parameter estimates:

MPG = 70.14 – 3.33 CYLINDERS + 0.02 Size – 0.20 HP – 0.3 WEIGHT – 0.78 ACCEL + 6.60 ENG\_TYPE

With a corrected total, or power, of 17 this provides us with a reduced sample size. This reduced sample size increases the amount of bias in this initial analysis. However, we will want to fill in the missing data to correct this reduction of sample size, which would in turn decrease the bias.

Before filling in the missing data, we determine the patterns of the missing data to establish what methods could be used to fill in the missing data.

| **Missing Data Patterns** | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **MPG** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **ACCEL** | **ENG\_TYPE** | **Freq** | **Percent** | **Group Means** | | | | | | |
| **MPG** | **CYLINDERS** | **SIZE** | **HP** | **WEIGHT** | **ACCEL** | **ENG\_TYPE** |
| **1** | X | X | X | X | X | X | X | 18 | 47.37 | 26.605556 | 5.333333 | 177.055556 | 101.888889 | 2.795333 | 14.355556 | 0.333333 |
| **2** | X | X | X | X | X | X | . | 2 | 5.26 | 31.350000 | 4.000000 | 95.000000 | 70.000000 | 2.125000 | 16.850000 | . |
| **3** | X | X | X | X | X | . | X | 1 | 2.63 | 18.200000 | 8.000000 | 318.000000 | 135.000000 | 3.830000 | . | 1.000000 |
| **4** | X | X | X | X | X | . | . | 1 | 2.63 | 17.600000 | 8.000000 | 302.000000 | 129.000000 | 3.725000 | . | . |
| **5** | X | X | X | X | . | X | X | 3 | 7.89 | 28.133333 | 4.666667 | 128.000000 | 72.666667 | . | 16.166667 | 0 |
| **6** | X | X | X | X | . | . | X | 1 | 2.63 | 21.500000 | 4.000000 | 121.000000 | 110.000000 | . | . | 0 |
| **7** | X | X | X | . | X | X | X | 5 | 13.16 | 22.320000 | 5.400000 | 182.800000 | . | 3.009800 | 15.240000 | 0.400000 |
| **8** | X | X | . | X | X | X | X | 2 | 5.26 | 19.100000 | 6.000000 | . | 115.000000 | 3.112500 | 15.150000 | 0 |
| **9** | X | X | . | X | . | X | X | 1 | 2.63 | 30.500000 | 4.000000 | . | 78.000000 | . | 14.100000 | 0 |
| **10** | X | . | X | X | X | X | X | 2 | 5.26 | 21.100000 | . | 176.000000 | 110.000000 | 3.087500 | 15.750000 | 0 |
| **11** | X | . | X | X | X | . | X | 1 | 2.63 | 18.100000 | . | 258.000000 | 120.000000 | 3.410000 | . | 0 |
| **12** | X | . | X | X | . | X | X | 1 | 2.63 | 17.000000 | . | 305.000000 | 130.000000 | . | 15.400000 | 1.000000 |

The data is not in any discernable pattern and can be considered arbitrary. Since the pattern of the missing data is considered arbitrary, we can use the MCMC method, which is the default method in PROC MI.

We will now create 5 separate single chain imputations from the data to fill in the missing data. Afterwards, we will combine the datasets. Each of the 5 imputations used all 38 of the observations that were read compared to our initial dataset that did not. Also, all of the imputations are considered more accurate as they all have a higher adjusted R squared value than our initial analysis.

The REG Procedure

Model: MODEL1

Dependent Variable: MPG

Imputation Number=1

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 38 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1475.92029 | 245.98672 | 69.22 | <.0001 |
| **Error** | 31 | 110.17050 | 3.55389 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.88518 | **R-Square** | 0.9305 |
| **Dependent Mean** | 24.76053 | **Adj R-Sq** | 0.9171 |
| **Coeff Var** | 7.61363 |  |  |

The REG Procedure

Model: MODEL1

Dependent Variable: MPG

Imputation Number=2

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 38 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1480.23596 | 246.70599 | 72.25 | <.0001 |
| **Error** | 31 | 105.85483 | 3.41467 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.84788 | **R-Square** | 0.9333 |
| **Dependent Mean** | 24.76053 | **Adj R-Sq** | 0.9203 |
| **Coeff Var** | 7.46302 |  |  |

The REG Procedure

Model: MODEL1

Dependent Variable: MPG

Imputation Number=3

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 38 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1478.18064 | 246.36344 | 70.77 | <.0001 |
| **Error** | 31 | 107.91015 | 3.48097 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.86574 | **R-Square** | 0.9320 |
| **Dependent Mean** | 24.76053 | **Adj R-Sq** | 0.9188 |
| **Coeff Var** | 7.53512 |  |  |

The REG Procedure

Model: MODEL1

Dependent Variable: MPG

Imputation Number=4

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 38 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1472.16820 | 245.36137 | 66.77 | <.0001 |
| **Error** | 31 | 113.92259 | 3.67492 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.91701 | **R-Square** | 0.9282 |
| **Dependent Mean** | 24.76053 | **Adj R-Sq** | 0.9143 |
| **Coeff Var** | 7.74220 |  |  |

The REG Procedure

Model: MODEL1

Dependent Variable: MPG

Imputation Number=5

|  |  |
| --- | --- |
| **Number of Observations Read** | 38 |
| **Number of Observations Used** | 38 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1460.37396 | 243.39566 | 60.02 | <.0001 |
| **Error** | 31 | 125.71683 | 4.05538 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2.01380 | **R-Square** | 0.9207 |
| **Dependent Mean** | 24.76053 | **Adj R-Sq** | 0.9054 |
| **Coeff Var** | 8.13310 |  |  |

Each of the models have different parameter estimates. However, this is to be expected because each of the models have different imputed values which will then create different parameter estimates. Now we can run analysis on all of the imputed data by combining it using PROC MIANALYZE.

| **Variance Information (5 Imputations)** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Variance** | | | **DF** | **Relative Increase in Variance** | **Fraction Missing Information** | **Relative Efficiency** |
| **Between** | **Within** | **Total** |
| **CYLINDERS** | 0.015726 | 0.568977 | 0.587847 | 3881.6 | 0.033166 | 0.032600 | 0.993522 |
| **SIZE** | 0.000112 | 0.000335 | 0.000469 | 48.531 | 0.402702 | 0.314759 | 0.940776 |
| **HP** | 0.000533 | 0.001484 | 0.002124 | 44.079 | 0.431107 | 0.330924 | 0.937924 |
| **WEIGHT** | 5.176341 | 7.777031 | 13.988640 | 20.286 | 0.798712 | 0.491796 | 0.910449 |
| **ACCEL** | 0.079946 | 0.072036 | 0.167971 | 12.262 | 1.331762 | 0.627338 | 0.888520 |
| **ENG\_TYPE** | 0.197465 | 2.258081 | 2.495039 | 443.48 | 0.104938 | 0.099026 | 0.980579 |
| **Intercept** | 4.645655 | 16.292645 | 21.867431 | 61.546 | 0.342166 | 0.278022 | 0.947325 |

| **Parameter Estimates (5 Imputations)** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Std Error** | **95% Confidence Limits** | | **DF** | **Minimum** | **Maximum** | **Theta0** | **t for H0: Parameter=Theta0** | **Pr > |t|** |
| **CYLINDERS** | -2.892369 | 0.766712 | -4.3956 | -1.38917 | 3881.6 | -3.037374 | -2.698869 | 0 | -3.77 | 0.0002 |
| **SIZE** | 0.030931 | 0.021663 | -0.0126 | 0.07447 | 48.531 | 0.020523 | 0.043585 | 0 | 1.43 | 0.1597 |
| **HP** | -0.158924 | 0.046085 | -0.2518 | -0.06605 | 44.079 | -0.197652 | -0.136972 | 0 | -3.45 | 0.0013 |
| **WEIGHT** | -3.214728 | 3.740139 | -11.0095 | 4.58001 | 20.286 | -6.129836 | -0.268453 | 0 | -0.86 | 0.4001 |
| **ACCEL** | -0.721966 | 0.409842 | -1.6128 | 0.16889 | 12.262 | -1.071906 | -0.352546 | 0 | -1.76 | 0.1030 |
| **ENG\_TYPE** | 5.855301 | 1.579569 | 2.7509 | 8.95967 | 443.48 | 5.192451 | 6.299408 | 0 | 3.71 | 0.0002 |
| **Intercept** | 69.542738 | 4.676262 | 60.1937 | 78.89183 | 61.546 | 67.011662 | 72.391093 | 0 | 14.87 | <.0001 |

The first chart above shows the difference between the values that were put in place of the missing data by each of the 5 imputations. The second chart details the parameter estimates for the 5 imputations. This gives us the following final equation via regression for calculating miles per gallon:

MPG = 69.54 -2.89 CYLINDERS + 0.03 SIZE -0.1 HP -3.21 WEIGHT – 0.72 ACCEL + 5.86 ENG\_TYPE.

According to this equation, the most impactful variables for determining the miles per gallon of each vehicle are the engine type, weight, and number of cylinders.

**Future Work**

When performing further analysis, we can compare how successful the multiple imputation method would be if we compared it to other analyses that used a different number of imputations. Also, we can compare these results to different methods of dealing with missing data such as last value carried forward (LUCF), hot desk imputation, or propensity score substitution.

Multiple imputation is an excellent tool for dealing with missing values and was able to provide us with a more accurate analysis of our data when performing regression in terms of bias and adjusted R squared.

**Appendix**

**data** car;

infile '\\Client\C$\Users\Phillip\Downloads\carmpgdata\_2\_2.txt' firstobs=**2** delimiter='09'x DSD;

input AUTO :$23. MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

ods html body = "\\Client\C$\Users\Phillip\Downloads\SASImputationOutput.doc";

**proc** **print** data=car;**run**;

ods graphics on;

title 'Predicting MGP';

**proc** **reg** data=car;

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

**run**;**quit**;

title;

ods graphics off;

title 'Examine Missing Pattern';

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

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

title;

title 'Create Series of Files with Imputed Data';

**proc** **mi** data=car out=miout nimpute=**5** seed=**35399**;

var MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**run**;

title;

title 'Run Analysis Using Imputed Data';

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

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

by \_Imputation\_;

**run**;

title;

title 'Combining Analyses Using PROC MIANALYZE';

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

modeleffects CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE Intercept;

**run**;

title;

ods html close;