## **Abstract**

A consistent problem for data analysts is missing information. Often discovered during the exploratory phase of data analysis, missing data is often responsible for a reduction in power, an increase bias and misrepresentation of outcomes in analysis tasks. Multiple imputation methods have become more applicable within the previous decade and provide an invaluable tool for data analysts dealing with missing data. In this paper, we investigate and compare results of fitting a multiple linear regression model after imputing missing data. Listwise deletion, a classical imputation method, is shown to produce different, less precise results in a multiple regression setting when compared to multiple imputation using a Bayesian simulation approach.

## **Introduction**

Statistical analysis is dependent on the quality of data for robust and accurate outcomes. While an ideal case would include accurate measurements for all variables in a dataset, in real world studies this is not always possible. Missing data can take on three major forms. Data can be missing completely at random (MCAR) - where no relationship exists between two variables in a dataset. Thus, missing data is completely random. When missing data is related to the outcome of the experiment or the response, the data is often labeled “Missing Not at Random” (MNAR). A third case, Missing at Random (MAR), occurs when a variable is found to be missing randomly, but only after controlling for a second variable’s state or value.

Typically, missing data is imputed in order to provide for a complete dataset, especially if the data is MAR or MCAR. We explore two methods, listwise deletion and multiple imputation. Listwise deletion removes all observations missing data, essentially reducing a dataset with missing observations to a complete dataset containing only observations with every variable represented. On the other hand, multiple imputation methods fill in missing data with feasible values based on the variability and uncertainty in the dataset, given observed data (1)**.**

In this work, we examine a dataset of automobiles with missing data for one or more features in over half of the observations. Our objective is to analyze the effects of two imputation methods on analysis outcomes. For each dataset produced by imputation methods, a multiple regression model is fit to predict miles per gallon, given explanatory variables discussed in a future section. These fits are compared to investigate differences driven by the two imputation methods implemented.

## **Literature Review**

Missing Data, and methods to correct for missing data, was the foundation of Little and Rubin’s work in (2). This often cited work outlines a researcher’s approach to missing data, starting with cataloging the pattern of missingness within the data. The authors provide some of the basic methods of correcting missing data - for example, simple substitution and outline some of the problems introduced with this method. The method fails to allow for the natural variability that exists in the remainder of the data set. The work also explains the method of multiple imputation, through which multiple sets of data are synthesized based on the distributions of existing variables, and run through the complete modeling process to complete multiple models, which are then averaged.

Stern in (3) helps explain the implications of missing data, and why data may be missing in case work. This paper has a very specific explanation of missing data types, and why they are important in a medical paper. He explains Missing Completely at Random” as data that might be lost due to a breakdown in a measurement device. Missing at Random might be related to the age of the subjects in a blood pressure study, because young people are less likely to have high blood pressure and practitioners less likely to regularly measure their blood pressure. The third state, Missing Not at Random, is explained as persons with high blood pressure are more likely to miss an appointment with a practitioner, because they may be experiencing a headache - which is more likely among persons with high blood pressure. Stern also addresses basic imputation methods, like last value carried forward, and addresses when this may be a pitfall in research. He also describes a study relating cholesterol and heart disease, published in the *British Medical Journal*, that had a changed outcome between considering only complete-data cases, and partial data cases.

## **Background**

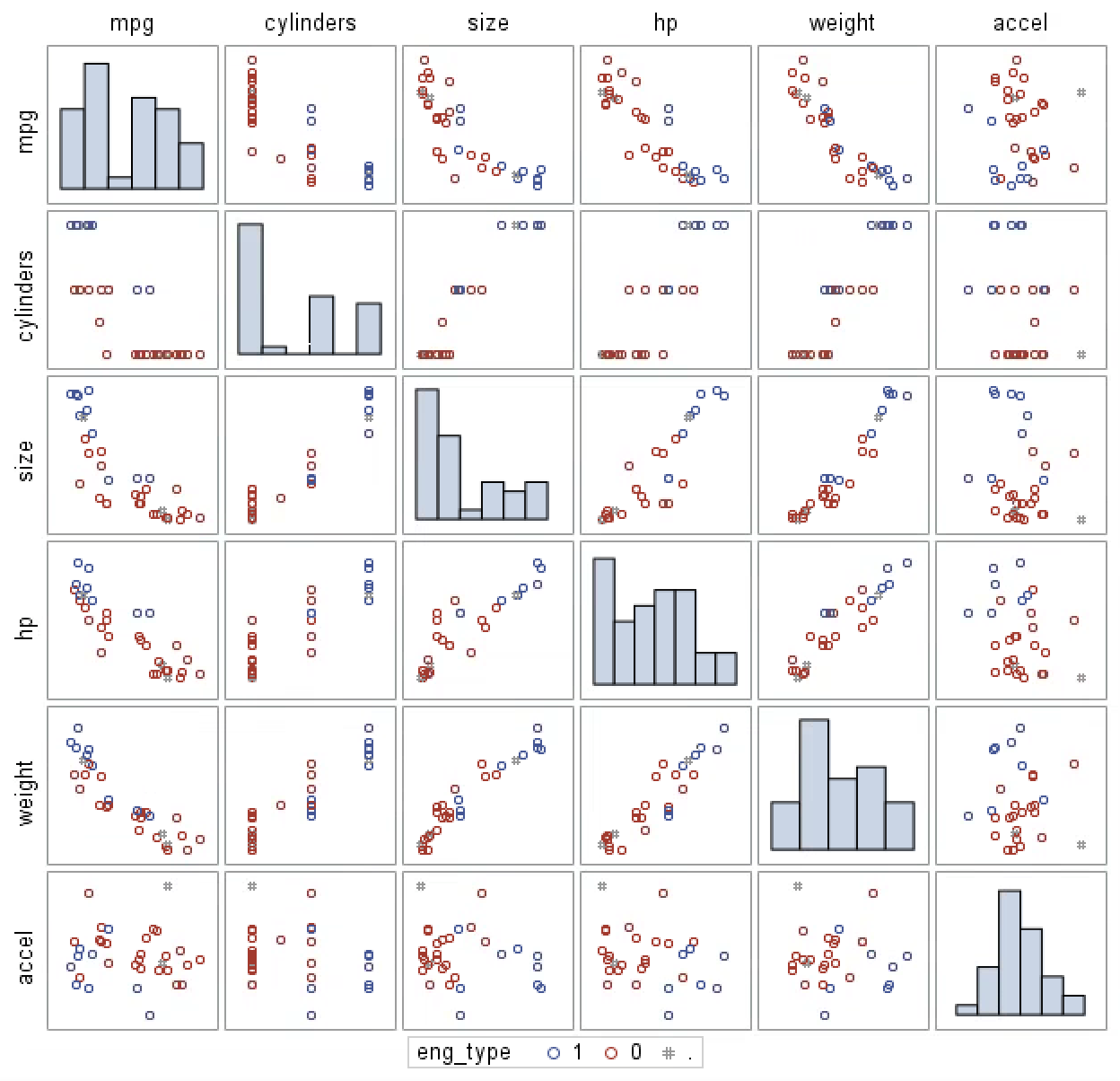
The automobile dataset contains 38 observations and eight variables. Of these 38 observations, 20 observations are missing data for one or more variables. Our objective is to analyze the effect of listwise deletion and multiple imputation methods on regression outcomes analyzing miles per gallon (fuel economy) as the response variable. A full list of variable descriptions, including the response, can be found in Table 1:

Table 1: Variable Descriptions and Types

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| Auto | Make and model of car (name) | String |
| MPG | Estimated miles per gallon | Continuous |
| Cylinders | Number of engine cylinders | Discrete |
| Size | Engine displacement measurement | Continuous |
| HP | Engine horsepower | Continuous |
| Weight | Weight of automobile | Continuous |
| Accel | Acceleration | Continuous |
| Eng\_Type | Engine type | Discrete |

Initial analysis of the raw data reveals strong linear correlations visually as well as via a correlation matrix. Of interest is the eng\_type variable which indicates “large” or “small” engines. A large engine, on average, has a size value of 281.9, while smaller engine sizes are 137.27. Other linear relationships between variables, such as weight and miles per gallon can be inferred quite easily from a simple scatterplot matrix as presented in Figure 1, particularly when broken down by engine type:

Figure 1: Scatterplot Matrix for Cars - Raw Data (SAS 9.4 Output)



Continuous explanatory variables, except for miles per gallon and size, show mostly normal distributions. Per guidance, no transformations or interactions are undertaken in this analysis. Because of such strong correlations, such as a 0.95, significant correlation between cylinders and size, it can also be argued our model may be susceptible to collinearity issues. However, given guidance, we maintain all variables in our analysis instead of investigating variable importance and pruning redundant explanatory variables. Additionally, the auto variable, giving the string name of each vehicle is unique for each record and unimportant in our analysis. Thus, we exclude the auto variable as an explanatory variable and move forward with six explanatory variables and one response variable, miles per gallon.

**Methods**

After exploration of the raw data, listwise deletion, a classical imputation method, is implemented using SAS. Listwise deletion removes any observation with missing data from the dataset. This means, if the observation is only missing horsepower (HP), it will be completely removed from the dataset. By default, SAS applies listwise deletion when running a multiple regression analysis unless otherwise specified.

Utilizing explanatory variables CYLINDERS, HP, SIZE, WEIGHT, ACCEL and ENG\_TYPE, a multiple regression model is fit to predict MPG (miles per gallon). This model automatically removes observations with missing values. Given over half of observations are deleted and our dataset contains less than 40 observations, significant power is lost. Additionally, we can expect parameters to be significantly impacted by removing over half of the observations. Per guidance, transformations and interactions are not implemented to take care of non-normality or explore additional effects. The raw data is used to fit the model with SAS PROC REG. Regression estimates are evaluated in the results section.

In the case of multiple imputation, additional methods must be considered in order to properly impute missing data. Simple listwise deletion does not require pre-work to ensure proper implementation. On the other hand, to ensure multiple imputation is implemented appropriately, we must analyze the missing data pattern in the dataset. Fortunately, SAS provides effective methods to analyze missing data patterns via PROC MI.

Methods of multiple imputation are dependent on the missingness pattern in the data. Further, we make an assumption that data are missing at random. Multiple imputation is often robust to assumption deviations if the data are missing at random or missing completely at random. Given that the pattern of data are visually random and no pattern is discernible when it comes to the outcome of miles per gallon, we can safely move forward with multiple imputation.

Additionally, implementations of multiple imputation should be selected based on the missingness pattern in the data given in Table 2. Monotonic patterns are often more well suited with multiple imputation based on specific data types. For instance, a continuous variable imputation would do well to use a linear regression-based method, while a categorical variable often uses logistic regression.

In our case, the missingness of the data shows an arbitrary pattern. Given this pattern, a Bayesian simulation method called Monte Carlo Markov Chain (MCMC) is used for multiple imputation. MCMC provides for a posterior distribution based on Markov chain random walks given the variability and uncertainty in the data, often via initial estimates of a covariance matrix and other critical parameters. This distribution stabilizes when there is no autocorrelation and is used to sample from in order to provide for imputed values to fill in missing data.

Table 2: Missing Data Patterns and Group Means (Means Bolded)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Missing Data Patterns | | | | | | | | | |
| Grp | MPG | Cyl | Size | HP | Weight | Accel | Eng\_Type | Freq | Pct |
| 1  Mean | X  **26.61** | X  **5.33** | X  **177.01** | X  **101.89** | X  **2.80** | X  **14.36** | X  **0.33** | 18 | 47.4 |
| 2 | X  **31.35** | X  **4.00** | X  **95.00** | X  **70.00** | X  **2.13** | X  **16.85** | - | 2 | 5.3 |
| 3 | X  **18.20** | X  **8.00** | X  **318.00** | X  **135.00** | X  **3.83** | - | X  **1.00** | 1 | 2.6 |
| 4 | X  **17.60** | X  **8.00** | X  **302.00** | X  **129.00** | X  **3.73** | - | - | 1 | 2.6 |
| 5 | X  **28.13** | X  **4.67** | X  **128.00** | X  **72.67** | - | X  **16.17** | X  **0.00** | 3 | 7.9 |
| 6 | X  **21.50** | X  **4.00** | X  **121.00** | X  **110.00** | - | - | X  **0.00** | 1 | 2.6 |
| 7 | X  **22.32** | X  **5.40** | X  **182.80** | - | X  **3.01** | X  **15.24** | X  **0.40** | 5 | 13.2 |
| 8 | X  **19.10** | X  **6.00** | - | X  **115.00** | X  **3.11** | X  **15.15** | X  **0.00** | 2 | 5.3 |
| 9 | X  **30.50** | X  **4.00** | - | X  **78.00** | - | X  **14.10** | X  **0.00** | 1 | 2.6 |
| 10 | X  **21.10** | - | X  **176.00** | X  **110.00** | X  **3.09** | X  **15.75** | X  **0.00** | 2 | 5.3 |
| 11 | X  **18.10** | - | X  **258.00** | X  **120.00** | X  **3.41** | - | X  **0.00** | 1 | 2.6 |
| 12 | X  **17.00** | - | X  **305.00** | X  **130.00** | - | X  **15.40** | X  **1.00** | 1 | 2.6 |

In our case, MCMC provides for five different imputed datasets, versus the single imputed dataset listwise deletion produces. These datasets provide for variability and uncertainty that listwise deletion often does not. Thus, we do not introduce as much bias in estimates or affect parameters of the data as drastically using multiple imputation. Further, these datasets provide for more power given a larger degrees of freedom since we’ve created complete datasets with all observations intact. This greater degrees of freedom allows us to gain more power in our analysis when compared to listwise deletion.

All of these datasets are subsequently combined and a multiple regression analysis is executed. The parameters and coefficients of the regression model are consolidated according to Rubin’s rules. Essentially, averaging coefficients while taking into account their variability (1). This is done in SAS via PROC MIANALYZE.

We compare the resulting regression models from listwise deletion and multiple imputation using MCMC to determine effects on coefficients and the overall model, including goodness of fit.

## **Results**

Comparing mean and standard errors of each imputation method, it is noticeable that the standard error for size and horsepower are different. As one would expect, multiple imputation provides additional data points according to the characteristics of the observed values in the dataset. These additional observations typically will reduce the variability parameters of features missing significant amounts of data. This plays out in Table 3.

Table 3: Parameter Comparison for Imputation Methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Listwise Deletion** | | **Multiple Imputation** | |
| Variable | Mean | Std Error | Mean | Std Error |
| Cylinders | 5.32 | 0.28 | 5.39 | 0.26 |
| Size | 180.89 | 15.45 | 180.44 | 14.52 |
| HP | 101.33 | 4.72 | 101.98 | 4.29 |
| Weight | 2.91 | 0.13 | 2.86 | 0.12 |
| Accel | 14.94 | 0.27 | 14.93 | 0.26 |
| Eng\_Type | 0.29 | 0.08 | 0.27 | 0.08 |

One consideration to be cognizant of is that PROC MI, unless told otherwise, imputes discrete values into continuous values. A clear case can be seen in the Buick Century Special, which is missing cylinders data.

Table 4: MCMC Results for Discrete Variable Cylinders

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Imputation** | **mpg** | **cylinders** | **size** | **hp** | **weight** | **accel** | **eng\_type** |
| 1 | 20.6 | 4.84 | 231 | 105 | 3.38 | 15.8 | 0 |
| 2 | 20.6 | 6.23 | 231 | 105 | 3.38 | 15.8 | 0 |
| 3 | 20.6 | 5.76 | 231 | 105 | 3.38 | 15.8 | 0 |
| 4 | 20.6 | 5.6 | 231 | 105 | 3.38 | 15.8 | 0 |
| 5 | 20.6 | .549 | 231 | 105 | 3.38 | 15.8 | 0 |

This results in unrealistic imputed values. For example, cylinders are integers between four and twelve. This can be expected when using MCMC as the PROC MI method in SAS expects continuous values. As this is a very linear dataset, and cylinders are strongly negatively correlated with miles per gallon, we could simply round to the nearest logical value. However, linear rounding has been shown to perform poorly when estimating variable means after imputation (4). Further, discriminant and logistic methods are only applicable when data is monotonic. Therefore, we are comfortable leaving cylinders and discrete imputation as is after PROC MI is ran.

Using mpg, cylinders, size, hp, weight, accel and eng\_type as explanatory variables, we fit multiple regression models for each imputation type: listwise deletion and multiple imputation. By using PROC REG on the raw dataset for cars, listwise deletion is applied by default. Fitting a multiple regression model on the raw data results in the loss of 20 observations. Thus, only 18 complete observations are used when utilizing listwise deletion for imputation. Model results from these 18 observations are found in Table 5.

Table 5: Multiple Linear Regression Results with Listwise Deleted Imputation

|  |  |  |
| --- | --- | --- |
| **Model Results** | | |
|  | **Value** | **Pr > F** |
| F | 22.34 | <0.0001 |
| Root MSE | 2.40 |  |
| Adj R-SQ | 0.88 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Regression Coefficients** | | | |
|  | **Estimate** | **Error** | **PR >|t|** |
| Intercept | 70.14 | 8.04 | <0.0001 |
| cylinders | -3.33 | 1.56 | 0.056 |
| size | 0.02 | 0.03 | 0.49 |
| hp | -0.20 | 0.08 | 0.03 |
| weight | -0.31 | 5.13 | 0.95 |
| accel | -0.78 | 0.58 | 0.21 |
| eng\_type | 6.60 | 3.59 | 0.09 |

Listwise imputation results in a significant regression model (F = 22.34) with an adjusted r-squared value of 0.88 and a root MSE of 2.40. Thus, even with only 18 observations, our model explains a significant amount of variance in the dataset. However, we see collinearity rear its head in the form of insignificant variables size (0.49) and weight (0.95). This is expected, as size and weight measurements are quite similar. Indeed, when weight is removed from the model, adjusted r-squared for the model is 0.89, indicating collinearity.

In order to compare the effectiveness of multiple imputation to listwise deletion, five imputed datasets are created using PROC MI and MCMC. For each imputed dataset, a regression model is fit on cylinders, size, horsepower, weight, acceleration and engine type for the response variable miles per gallon. These regression models are combined to form consolidated regression coefficients. For clarity a side by side coefficient charts for listwise and multiple imputation is found in Table 6.

Table 6: Regression Coefficients for Listwise Deletion and Multiple Imputation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regression Coefficients: Listwise Deletion** | | | | | | |
| **Variable** | **Estimate** | | **Standard Error** | | **PR >|t|** | |
| **Imputation Type** | Multiple | Listwise | Multiple | Listwise | Multiple | Listwise |
| Intercept | 66.71 | 70.14 | 5.22 | 8.04 | < 0.0001 | < 0.0001 |
| Cylinders | -2.04 | -3.33 | 0.98 | 1.56 | 0.04 | 0.06 |
| Size | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.49 |
| HP | -0.13 | -0.20 | 0.05 | 0.08 | 0.01 | 0.03 |
| Weight | -6.42 | -0.31 | 3.51 | 5.13 | 0.07 | 0.95 |
| Accel | -0.53 | -0.78 | 0.40 | 0.58 | 0.19 | 0.21 |
| Eng Type | 3.54 | 6.60 | 1.87 | 3.59 | 0.06 | 0.09 |

Of particular note are the large differences in weight and engine type variables. Weight, in particular, becomes nearly significant at p=0.07 in the regression model after multiple imputation. Further, it is expected that given additional observations based on the natural variability in the dataset, we would see less volatile estimates. This is proven out in the multiple imputation scenario, where each explanatory variable experiences an improvement in standard error. SAS conveniently provides relative variance increases, which explains the additional variability the model incurs given that data is missing. Size and weight experience variability improvement, given nearly 30 percent of observations are missing from these variables.

The regression model based on multiple imputation also performs slightly better than the regression model based on listwise deletion. With an average root mean square error of 2.33, the multiple imputation-based regression model’s performance is better than the listwise deletion version, which obtained a root mean square error of 2.40.

Table 7: Error and Coefficient Variation Among Imputation Runs

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Run** | **RMSE** | **Intercept** | **cylinders** | **size** | **hp** | **weight** | **accel** | **eng\_type** |
| 1 | 2.15 | 67.36 | -1.77 | 0.05 | -0.11 | -8.06 | -0.54 | 3.03 |
| 2 | 2.36 | 66.17 | -2.27 | 0.05 | -0.11 | -7.32 | -0.43 | 3.88 |
| 3 | 2.53 | 67.16 | -1.97 | 0.05 | -0.15 | -6.51 | -0.53 | 2.85 |
| 4 | 2.14 | 65.72 | -2.54 | 0.04 | -0.12 | -6.53 | -0.35 | 4.67 |
| 5 | 2.50 | 67.14 | -1.63 | 0.02 | -0.16 | -3.67 | -0.80 | 3.28 |
| Mean | 2.33 | 66.71 | -2.04 | 0.04 | -0.13 | -6.42 | -0.53 | 3.54 |

## **Future Work, Discussion Conclusions, and Next Steps**

There is an obvious case for multiple imputation to complete the observations of a data study. In this example, we completed a small set of observations of vehicle traits to attempt to better forecast vehicle fuel efficiency. In this case, half of the observations had some degree of missingness. The results provided by SAS’ PROC MI and MIANALYZE were largely similar across the individual imputations and the resulting linear regressions. This offers a level of confidence to a researcher implementing this approach for real work.

The United States Department of Transportation implemented the Corporate Average Fuel Economy (CAFE) standard for the passenger vehicle industry in the United States following several of the oil crises in the early 1970’s. These standards set a required minimum fuel economy across a manufacturer’s entire fleet, weighted on sales of individual models. It could be imagined that an analyst for a car manufacturer, attempting to provide input to the company’s future strategy and vehicle lineup, might use a data set that includes this data, but might be missing some measurements of some variables. Multiple imputation would allow the analyst to use a larger amount of the data set, and derive a less biased viewpoint of the relationship between the variables that define a car and its fuel economy.

The SAS implementation of Multiple Implication offers a Markov Chain Monte Carlo model, as well as a fully conditional specification variant, that uses a joint distribution for all variables. As data science practitioners, the idea of multiple imputation is an important concept that needs to be considered following exploratory data analysis, no matter the analysis or machine learning platform selected. The R computing language offers several libraries that provide multiple imputation based on different algorithms, including a random forest method. The Python language, with its pandas data structure, offers simple imputation methods. The fancyimpute library offers a chained equation implementation for imputation of missing values.

Application of missing data imputation for machine learning applications has been less documented in academic papers. In some applications, it is likely being overcome with the size of the data set, or the measurement methods otherwise preclude incomplete measurements in the process.

## References

1. Yuan, Y. C.. “Multiple imputation for missing data: Concepts and New Developments (Version 9.0).” SAS Institute Inc, Rockville, MD, 49, 1-11.
2. Little, Roderick JA, and Donald B. Rubin. “Statistical Analysis with Missing Data.” John Wiley & Sons, 2014.
3. Sterne, Jonathan AC, et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *Bmj* 338 (2009): b2393.
4. Allison, Paul. “Imputation of Categorical Variables with PROC MI.” Paper 113-30. University of Pennsylvania, 2014.

## Appendix - SAS Code

data carmpg;

infile '\\client\c$\Users\patrickcorynichols\MSDS\_QTW\CaseStudy2\_MSDS7333\carmpgdata\_2\_2\_2.txt'

DSD delimiter='09'x DSD missover firstobs = 2;

length auto $25;

input auto mpg cylinders size hp weight accel eng\_type;

RUN;

PROC PRINT data = carmpg;

RUN;

/\* Begin exploratory data analysis \*/

PROC MEANS data = carmpg N NMISS MEAN STD STDERR CLM Q1 MEDIAN Q3;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

RUN;

PROC MEANS data = carmpg N NMISS MEAN STD STDERR CLM Q1 MEDIAN Q3;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL;

CLASS ENG\_TYPE;

RUN;

PROC SGSCATTER data = carmpg;

title 'SP Matrix for Cars';

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

RUN;

PROC CORR data = carmpg;

RUN;

Title 'Predicting MPG with Listwise Deletion Imputation';

PROC REG data = carmpg;

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

OUTPUT OUT = modeldiags

RSTUDENT = rstudent

COOKD = cookd;

RUN;

QUIT;

Title 'Identifying Missing Patterns with PROC MI';

ODS SELECT MISSPATTERN;

PROC MI data = carmpg seed = 100 nimpute = 0;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

MCMC;

RUN;

QUIT;

TITLE 'Create Imputed Datasets with Multiple Imputation';

PROC MI DATA = carmpg OUT = miout SEED = 100 NIMPUTE = 5;

VAR CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

MCMC;

RUN;

QUIT;

PROC PRINT data = miout;

RUN;

/\* Take a look at effects by automobile description for each imputation \*/

PROC SORT data = miout;

BY auto;

RUN;

PROC PRINT data = miout;

BY auto;

RUN;

TITLE 'MLR on Each Imputed Dataset';

PROC REG DATA = miout OUTEST = outreg COVOUT;

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

BY \_IMPUTATION\_;

RUN;

QUIT;

Title 'Combining Parameters According to Rubins Rules';

/\* analyze outputs vs listwise \*/

PROC MIANALYZE DATA = outreg;

MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE Intercept;

RUN;