**Abstract**

Missing data is often part of any analysis performed by data analysts. Typically, missing data is discovered during the exploratory phase of data analysis, often called exploratory data analysis (EDA). As a result of missing data, the output of the analyst’s work could be biased, outcomes misrepresented, and the power of the overall analysis reduced. In this paper, we analyze and compare results of fitting a multiple linear regression model after imputing missing data. Multiple imputation using a Bayesian simulation approach shows a more precise result in multiple regression when compared to the commonly used method of listwise deletion.

**Introduction**

Many times, in statistical analysis an analyst is often faced with missing data. Missing data often comes in three major forms. Data can be missing completely at random (MCAR), this is when no relationship exists between two variables in a dataset. The missing data is completely random. At times the missing data is related to the outcome of the response, or experiment. In this situation the data is labeled “Missing Not at Random” (MNAR). The third case, Missing at Random (MAR), can occur when a variable is missing randomly, but only after controlling for the second variable’s state, or value.

Missing data is often imputed to handle missing values to create a complete dataset. This often occurs if the data is MAR or MCAR. We will use two methods, listwise deletion and multiple imputation. Listwise deletion simply removes all observations with missing data. This gives a complete dataset, but with missing the observations that had incomplete data. The second method known as multiple imputation fill in missing data with a regression model to fill in the missing data with multiple iterations of the process[1].

The motivation for this analysis is supplied by a dataset of automobiles with missing data. The data is missing for one or more features in over half the observations. The goal in analyzing this dataset is to observe the effects of the two imputation methods. In each dataset produced by the imputation method a multiple regression model is fit to predict miles per gallon (MPG), given the explanatory variables described in the background section of the paper. The fits of the imputation methods are then compared to uncover the difference.

**Literature Review**

Research on missing data often reflects to the work of Donald Rubin[1]. In Rubin’s 1976 paper “Inference and missing data”, Rubin outlines the concern of ignoring the pattern of missing data and how it may lead inference on an analysis to be incorrect[2]. The authors of these papers offer solutions to missing data by substitution. They also give both pros and cons of various methods in handling missing data.

In Braccini et. al.[3] the authors describe the effects of missing vessel data on the fishing of the Australian eastern king prawn. During years of complete data from vessels power increased. However, with missing vessel data considerable errors would be introduced with a changing fishing vessel fleet. The analyses ventures through methods of data imputation and the implications to the fishery of continued data capturing methods for the management of the fishery. Many vessel leaders failed to log upgrades to vessels. If gaps originated randomly the estimates were not greatly affected. It was found that missing vessel information is random. They found that skippers from less efficient vessels failed to fill out the logs to complete the vessel information. These less “diligent” skippers introduced considerable error into the metrics measured.

**Background**

The automobile dataset contains 38 observations and eight variables. Of the 38 observations 18 are missing one or more variables. The goal is to provide an analysis of the difference between listwise deletion and multiple imputation methods on regression for the MPG as our response variable. A full list of variables and their description can be found in Table 1:



Table 1 - Variable Descriptions and Types

During EDA of the data strong linear correlations exist both in terms of visualization and in terms of a correlation matrix. The variable Eng\_type which indicates “large” or “small” engines is of particular interest. A large engine has an average size of 281.9 cubic inches, while a small engine size average 137.27 cubic inches. A larger engine is also likely to result in a higher weight of the car. More linear relationships can be inferred from a scatterplot matrix shown in Figure 1. The below figure is broken down by engine type.

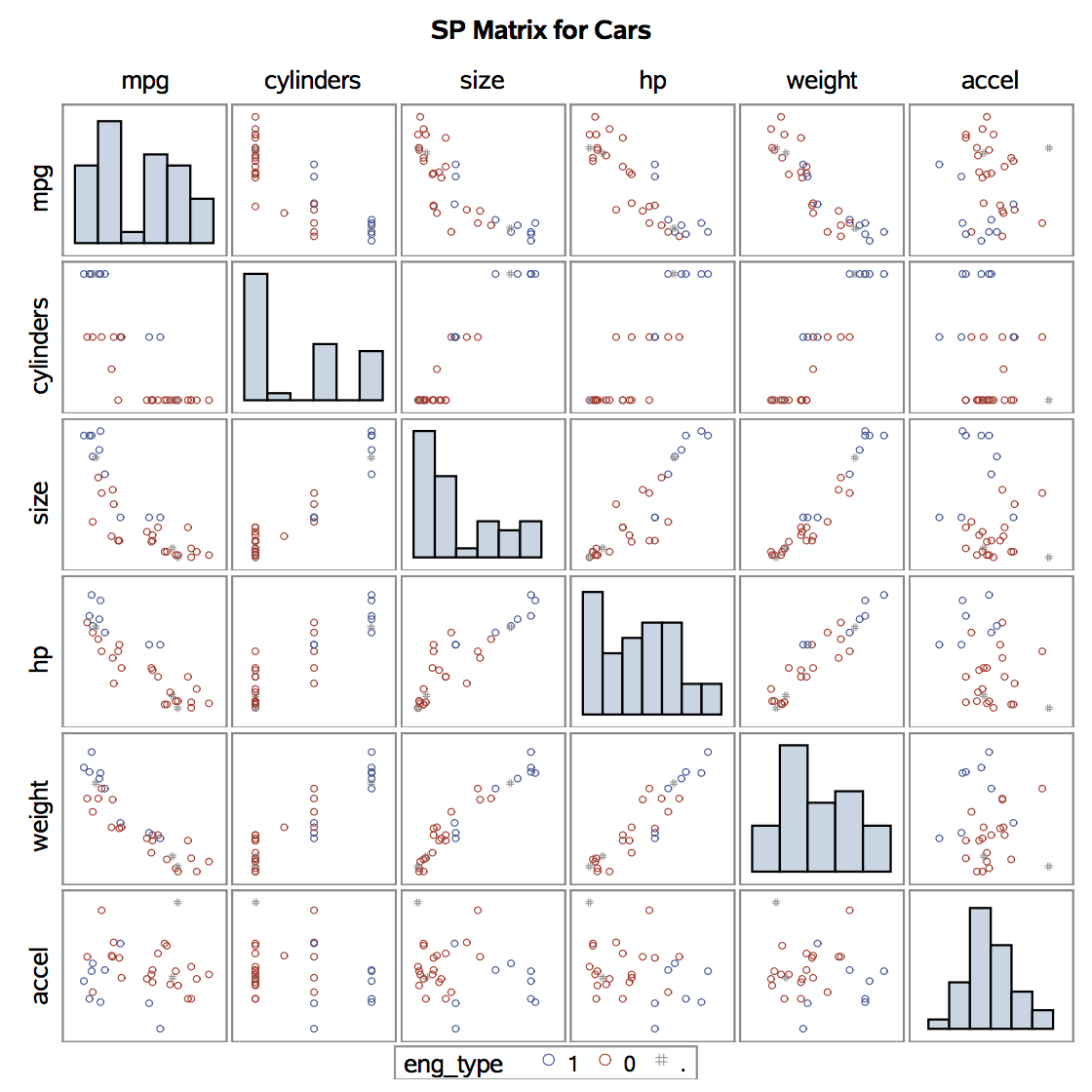


Figure 1 - Scatterplot Matrix for Cars – Raw Data (SAS Output)

Evidence for normal distributions can be found in most of the continuous variables with exception for MPG and size. Pearson correlation coefficients show strong relationships between weight of car and size of engine in cubic inches with 0.95 correlation there is likely sensitivity to collinearity issues. With that known we will proceed with guidance and maintain all variables in our analysis. The auto variable, which is a string name of each car is unique for each record and is unimportant to our analysis. We will exclude the auto variable as an explanatory variable. This will land our analysis with six explanatory variables and one response variable, MPG.

**Methods**

After the raw data set has been explored, listwise deletion will be implemented. Listwise deletion eliminates all entries with missing values for any of the attributes. SAS will use listwise deletion by default when running PROC REG. Over half of observations (20 of 38) are missing at least one attribute, so significant power is lost. Additionally, we can expect estimations 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, we must analyze the missing data pattern in the dataset. We will use SAS PROC MI to analyze the missing data patterns.

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 in Cars Dataset

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 degree of freedom since we’ve created complete datasets with all observations intact. This greater degree 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

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 an observation in our dataset missing cylinders data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Imputation** | **mpg** | **cylinders** | **size** | **hp** | **weight** | **accel** | **eng\_type** |
| 1 | 20.6 | 4.8 | 231.0 | 105.0 | 3.4 | 15.8 | 0.0 |
| 2 | 20.6 | 6.2 | 231.0 | 105.0 | 3.4 | 15.8 | 0.0 |
| 3 | 20.6 | 5.8 | 231.0 | 105.0 | 3.4 | 15.8 | 0.0 |
| 4 | 20.6 | 5.6 | 231.0 | 105.0 | 3.4 | 15.8 | 0.0 |
| 5 | 20.6 | 0.5 | 231.0 | 105.0 | 3.4 | 15.8 | 0.0 |

Table 4 - MCMC Results for Discrete Variable Cylinders

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

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.

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

Of note are the large differences in weight and engine type variables. Weight 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 Among Multiple Imputation Runs

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

The case for multiple imputation to complete a data set is solid. In the case above, over half of the observations contained missing data and would be excluded from the analysis using listwise deletion. Running multiple seeds of imputation and getting similar results drives confidence in this approach.

Something to look at going forward is octane requirements. The United States Department of Transportation Corporate Average Fuel Economy (CAFE) standards will require new vehicles to achieve over 50 mpg by 2025. In response, automakers have begun looking at higher octane fuel to drive higher compression ratios, resulting in greater fuel efficiency. Some refiners are exploring the economics of shifting the standard octane ratings from 87 (regular) and 91 (premium) to 95 (regular) and 101 (premium). Octane does not materially impact fuel economy, but engines designed for higher octanes can achieve greater efficiency.

One additional factor that will need to be considered in future studies is drivetrain. The rise of hybrid technology and fully electric drivetrains will drive average mpg up and will help break through the efficiency barriers of the traditional internal combustion engine. The correct approach for hybrid vehicles would be to stratify and determine if there are other metrics that may impact performance, such as battery size and composition.

**References**

1. Enders, C. K., “Multiple imputation as a flexible tool for missing data handling in clinical research”, Behavior Research and Therapy 90 (2017) 4-18
2. Rubin, D. B., “Inference and missing data”, Biometrics (1976)
3. Braccini, J.M. et. al., “Fishing power and standardized catch rates: implications of missing vessel-characteristic data from the Australian eastern king prawn (Melicertus plebejus) fishery, Canadian Journal of Fisheries and Aquatic Sciences (2012)

**Appendix - SAS Code**

/\*

Case Study 1 Code

Mike Crowder, Gerardo Garza,

Brian Kolovich and Brandon Lawrence

MSDS 7333

\*/

/\* Bring data in \*/

data carmpg;

infile '/home/mcrowder0/Crowder/QTW/carmpgdata\_26.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;

/\* Run proc means by Eng\_type to see effect \*/

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

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL;

CLASS ENG\_TYPE;

RUN;

/\* Scatterplot Matrix \*/

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;

/\*regression - listwise deletion\*/

title "Listwise Deletion";

proc reg data = carmpg;

model mpg = cylinders size hp weight accel eng\_type;

run;

/\*examine missing pattern\*/

title "Examining Missing Data";

ods select misspattern;

proc mi data = carmpg nimpute=0;

var mpg cylinders size hp weight accel eng\_type;

run;

/\*since missing data is arbitrary, use default method (MCMC)\*/

title "PROC MI MCMC Method - 1st Iteration";

proc mi data = carmpg out = miout seed = 113660;

var mpg cylinders size hp weight accel eng\_type;

run;

/\*run analysis using imputed data\*/

proc reg data = miout outest = outreg covout;

model mpg = cylinders size hp weight accel eng\_type;

by \_imputation\_;

run;