# Abstract

In this paper we take a look at missing data in a set of automobile stats and analyze the effect of the missing data on the statistical power of creating regression models from the data. The data is made of categorical data relating to the features of different car models and their fuel efficiency. We will review some potential shortcomings of using multiple imputation along with future considerations when imputing missing values. We initially run a regression analysis on the data with the fuel efficiency (miles per gallon) used as the response variable. After our initial results we ran the dataset through the statistical package SAS using the PROC MI function to impute all missing values. We concluded that imputing the missing values with PROC MI gave a more statistically significant result that what we started with.

# Introduction

Missing data is going to be a common issue in many datasets and depending on the nature of the dataset this missingness can be attributed to different reasons. Data can be collected from many different sources, which can cause issues depending on why the data is missing. Many statistical software packages including SAS will drop observations with any missing values. [1] This can be a problem as the power of any statistical tests can be reduced or bias can be introduced to the final model. For these reasons there needs to be an understanding of the type of missingness of the data and consideration of the best method of dealing with the missing values.

There are a number of basic methods of dealing with missing data from deleting observations with missing values to replacing the missing value with a substitute value. Substituting values can be simple replacement of the mean of the variable or imputing the value based on a regression based on other variables that make up the dataset. Imputing data can be of concern because in the end an assumption is being made regarding the missing data and the reason of the missingness. If the imputed values vary drastically that the true values of the missing data, then the resulting conclusion made from the imputed data can be misleading or erroneous.

For this analysis we will look at a small dataset of MPG data for different car models, along with other continuous explanatory variables, which are used to predict MPG. We will conduct an initial analysis with the missing data, using the default process used by SAS (listwise deletion), then we will conduct multiple imputation and note the difference in statistical power between the two results.

# Background

There are three primary types of missing data: missing completely at random (MCAR), missing not at random (MNAR), and missing at random (MAR). As explained by Donders and his associates, MCAR and MNAR missingness is going to be less common than MAR due to the specific reasons that MCAR and MNAR can occur. MCAR can be due to lost testing data, broken testing equipment, or other circumstances where the missing data is not dependent on the value of another variable. MNAR occurs when observations for a variable are missing depending on the value of the variable, with an example by Donders being that on a survey, values of income level may be missing for those at higher income levels may be due to the respondents’ reluctance to comment on their income. MAR occurs when data is that is missing can be correlated to the value of another variable, such as in the situation where healthcare data is collected, and fewer results for blood tests are collected for younger patients. With MAR, once this additional variable is taken in to account, the missingness of the data is easier to explain. [2]

When dealing with MAR data, most common techniques used for handling data can introduce bias. Due to simple techniques failing with MAR, single imputation or multiple imputation must be used to avoid bias in any final models. [2] Multiple imputation, which will be discussed later, can be a powerful tool in working to address datasets with missingness due to MAR. It should be noted that while multiple imputation can be powerful, it should be used with the direction of someone that is familiar with the pitfalls of the method as whatever results that are reported should be done with understanding of the underlying methods being performed by multiple imputation. As cited by Sterne and his associates, there have been published results that have been questioned due to the use of multiple imputation. [3]

There are four common ways to deal with missing data in data sets; listwise deletion, pairwise deletion, single imputation and multiple imputation. Each of these have different advantages and disadvantages discussed below.

When addressing missing data, one may be tempted to simply ignore any records where data is missing. These records are deleted from the data set and the analysis continues on the data where all records are complete. This is called listwise deletion or complete case analysis. Listwise deletion assumes the missing data is either missing at random or missing completely at random. If only a few pieces of data are missing, simply deleting them may be acceptable. Although in some cases this technique may produce valid results, there are some potential problems with listwise deletion. First is the reduction of the sample size of the data set. If the data set we are analyzing is not big to begin with, deleting a significant number of records reduces the power. Secondly, if too many records are deleted we may skew the data without any way to detect this. Also, some types of studies may inherently cause the same field in a record to be missing. Confidential or potentially embarrassing information revealed in a survey may be skipped by the people who don’t want to reveal some personal information where this field is filled in by people who are not concerned with providing the information. This may cause a bias in the data when the records with missing data are deleted.

Pairwise deletion is similar to listwise deletion in that it deletes missing data, but in the case of pairwise deletion, the data remaining in the record is retained to be used in the analysis. All the data available is used for analysis. Pairwise deletion assumes the missing data is missing completely at random. However, problems do crop up when using pairwise deletion. One problem is the statistics derived for one of the variables is based on a different sample size than those from another variable. As seen in Table 1, ***Var1***and ***Var4*** each have a sample size of 10, ***Var2*** has a sample size of 9 and ***Var3*** has a sample size of 8. The end result is each of these variables is using a different data set to derive any statistics.

Table – Pairwise Deletion Example

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observation | Var1 | Var2 | Var3 | Var4 |
| 1 | 0 | 75 | 135 | 12 |
| 2 | 1 | 63 | 158 | 11 |
| 3 | 1 | . | 192 | 15 |
| 4 | 0 | 85 | 145 | 20 |
| 5 | 1 | 58 | . | 12 |
| 6 | 0 | 92 | 138 | 13 |
| 7 | 1 | 82 | . | 18 |
| 8 | 1 | 71 | 129 | 17 |
| 9 | 0 | 68 | 170 | 15 |
| 10 | 0 | 80 | 161 | 16 |

Another result of different sample sizes for each variable is the correlation matrix that does not provide useful information. In the case above, any correlation between ***Var1*** and ***Var4*** which have the same sample size cannot be compared to any correlation between ***Var1*** and ***Var3*** since these two variables have different sample sizes.

Imputation is a method by which missing values in a data set are filled in. There are two classes of imputation available to the data scientist. These are single imputation and multiple imputation. Within single imputation several different approaches can be taken, with each approach have advantages and disadvantages. Single imputation fills in the missing data by finding a value that is likely to be close to the missing value. These values are used to run whatever analysis the data scientist is interested in. There are a number of single imputation methods available. One of the simplest methods is the mean substitution. The mean of all the available values for a variable is taken and then this value is substituted for all the missing values as seen in Table 2. This method results in underestimating the error because of changes in the correlation between variables. These problems are reasons not to use the mean substitution method.

Table - Mean Substitution Example

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observation | Var1 | Var2 | Var3 | Var4 |  | Var1 | Var2 | Var3 | Var3 |
| 1 | 0 | 75 | 135 | 12 |  | 0 | 75 | 135 | 12 |
| 2 | 1 | 63 | 158 | 11 |  | 1 | 63 | 158 | 11 |
| 3 | 1 | . | 192 | 15 |  | 1 | 67.3 | 192 | 15 |
| 4 | 0 | 85 | 145 | 20 |  | 0 | 85 | 145 | 20 |
| 5 | 1 | 58 | . | 12 |  | 1 | 58 | 153.5 | 12 |
| 6 | 0 | 92 | 138 | 13 |  | 0 | 92 | 138 | 13 |
| 7 | 1 | 82 | . | 18 |  | 1 | 82 | 153.5 | 18 |
| 8 | 1 | 71 | 129 | 17 |  | 1 | 71 | 129 | 17 |
| 9 | 0 | 68 | 170 | 15 |  | 0 | 68 | 170 | 15 |
| 10 | 0 | 80 | 161 | 16 |  | 0 | 80 | 161 | 16 |

A second method is called the dummy variable method. This method creates a new variable in the data set which tracks if a data field is missing or not. If the data is missing, the value of the dummy variable is set to 0 and set to a 1 if data is present. This method tracks where data is missing and may help the data scientist find patterns in the missing data. This method suffers from similar issues to the mean substitution method where error is reduced, and correlation is thrown off. An example is shown in Table 3 where ***Dummy1*** is the variable used to track the presence of ***Var2*** and ***Dummy2*** is the variable used to track the presence of ***Var3***.

Table - Dummy Variable example

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Observation | Var1 | Var2 | Var3 | Var4 | Dummy1 | Dummy2 |
| 1 | 0 | 75 | 135 | 12 | 1 | 1 |
| 2 | 1 | 63 | 158 | 11 | 1 | 1 |
| 3 | 1 | . | 192 | 15 | 0 | 1 |
| 4 | 0 | 85 | 145 | 20 | 1 | 1 |
| 5 | 1 | 58 | . | 12 | 1 | 1 |
| 6 | 0 | 92 | 138 | 13 | 1 | 0 |
| 7 | 1 | 82 | . | 18 | 1 | 1 |
| 8 | 1 | 71 | 129 | 17 | 1 | 0 |
| 9 | 0 | 68 | 170 | 15 | 1 | 1 |
| 10 | 0 | 80 | 161 | 16 | 1 | 1 |

Multiple imputation attempts to keep the underlying characteristics of the data set intact when replacing missing values. Multiple imputation works by creating copies of the original dataset and imputing possibly values of the missing data with taking values that follow a distribution that mirrors the underlying data. This method of creating values based on distributions for the variable is used because the true value of the missing value cannot be known, so all existing information is used to make an accurate estimation of the missing information. After these multiple datasets are created with sets of multiple imputed values then the datasets are plugged into the desired model, which are then averaged together to get the final model. [3]

Multiple imputation is useful because it allows the use of the original dataset, while also accounting for the variability of other information that is present, along with accounting for the effect of the imputed data on the final model. This method is not without its pitfalls since the imputed data can create a completely different model than if the data was not missing. This was referenced by Sterne and his associates in their analysis of the use multiple imputation for epidemiological studies and the creation of a tool to predict risks associated with cardiovascular problems. In the referenced study the authors found that there was no link between cardiovascular disease and cholesterol, after the use of multiple imputation to replace missing data. This result was later found to be counter to the results when all observations with missing data were removed, instead there was found to be a positive link between cardiovascular disease and the cholesterol. Even worse it was found that with an improved imputation method, a positive link was found between cholesterol and cardiovascular disease. [3]

Some issues with multiple imputation that should be considered. The outcome variable must be included when running the data through multiple imputation. If the outcome variable is not included when imputing missing variables, the association between any explanatory and response variables can be weakened. Most procedures for multiple imputation are on the assumption that the data is normally distributed, which can mean it can be difficult to use multiple imputation with binary or categorical data. Additionally, as mentioned earlier there have been issues found with some uses of multiple imputation for datasets where a large number of observations are missing for a given variable. [3]

# Method

Initially, to get an understanding of what our data looks like, we will perform an analysis on the data prior to performing multiple imputation so we can compare our results at the end. When running linear regression in SAS with missing data, SAS by default uses listwise deletion. [1] Table 4 confirms SAS is using listwise deletion by showing 18 values were not used in the analysis. The Analysis of variance table in Table 5 shows we have lower degrees of freedom leading to lower power. The parameter estimates are shown in Table 6.

Table - Number of missing values in original analysis

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

Table - Original Analysis of Variance

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 791.80042 | 158.36008 | 25.54 | <.0001 |
| **Error** | 14 | 86.81158 | 6.20083 |  |  |
| **Corrected Total** | 19 | 878.61200 |  |  |  |

Table - Original parameter estimates

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 67.61816 | 7.12819 | 9.49 | <.0001 |
| **CYLINDERS** | 1 | -1.19508 | 1.13851 | -1.05 | 0.3116 |
| **SIZE** | 1 | 0.05221 | 0.02938 | 1.78 | 0.0973 |
| **HP** | 1 | -0.15009 | 0.07848 | -1.91 | 0.0765 |
| **WEIGHT** | 1 | -6.71776 | 3.98252 | -1.69 | 0.1138 |
| **ACCEL** | 1 | -0.68451 | 0.44024 | -1.55 | 0.1423 |

The first step in imputing the data is to see if there are any patterns to the missingness of the data; is there a monotone pattern to missing data. Table 7 shows there is no discernable pattern to the missing data, so we will proceed as the missing data is non-monotone.

Table - Missing Data Pattern from Proc MI

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

Once we have confirmed the pattern of missingness is arbitrary, we use PROC MI in SAS to create the imputed data sets. Table 8 confirms we have performed 25 imputations

Table - Multiple Imputation Information

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.MPG |
| **Method** | MCMC |
| **Multiple Imputation Chain** | Single Chain |
| **Initial Estimates for MCMC** | EM Posterior Mode |
| **Start** | Starting Value |
| **Prior** | Jeffreys |
| **Number of Imputations** | 25 |
| **Number of Burn-in Iterations** | 200 |
| **Number of Iterations** | 100 |
| **Seed for random number generator** | 1234 |

Now that we have 25 resulting data sets, we will look at only the first one and compare it to what we originally saw in the regression analysis of the original data set using listwise deletion. Table 9 shows there are no longer any missing values in the data set and Table 10 - Imputed Data #1 Analysis of Variance shows we have 37 degrees of freedom allowing for more power. If we compare Table 6 - Original parameter estimates to Table 11 - Imputed Data #1 Parameter Estimates, there is some difference in the estimates, but the change is not dramatic.

Table - Imputed Data #1 Information

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

Table - Imputed Data #1 Analysis of Variance

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 1423.56369 | 284.71274 | 56.06 | <.0001 |
| **Error** | 32 | 162.52710 | 5.07897 |  |  |
| **Corrected Total** | 37 | 1586.09079 |  |  |  |

Table - Imputed Data #1 Parameter Estimates

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 67.24323 | 4.80907 | 13.98 | <.0001 |
| **CYLINDERS** | 1 | -1.38213 | 0.72490 | -1.91 | 0.0656 |
| **SIZE** | 1 | 0.06855 | 0.01741 | 3.94 | 0.0004 |
| **HP** | 1 | -0.10418 | 0.03928 | -2.65 | 0.0123 |
| **WEIGHT** | 1 | -10.09231 | 2.38944 | -4.22 | 0.0002 |
| **ACCEL** | 1 | -0.50312 | 0.30613 | -1.64 | 0.1101 |

The final step is to combine the results from the 25 imputed data sets. We run PROC MIANALYZE to perform this task. Table 12 shows the resulting parameter estimates after combining the 25 imputations.

Table - Results of combining imputations

| **Parameter Estimates (25 Imputations)** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Std Error** | **95% Confidence Limits** | | **DF** | **Minimum** | **Maximum** |
| **cylinders** | -0.998923 | 0.853992 | -2.6741 | 0.67627 | 1450.5 | -1.550900 | -0.362848 |
| **size** | 0.043529 | 0.024752 | -0.0051 | 0.09219 | 391.97 | 0.017026 | 0.068546 |
| **hp** | -0.117772 | 0.053495 | -0.2228 | -0.01273 | 664.59 | -0.165938 | -0.077031 |
| **weight** | -7.531299 | 3.641991 | -14.7021 | -0.36046 | 265.69 | -10.092315 | -4.067058 |
| **accel** | -0.700246 | 0.411786 | -1.5107 | 0.11020 | 291.6 | -1.035521 | -0.214607 |
| **Intercept** | 66.361776 | 5.942839 | 54.6938 | 78.02979 | 697.04 | 59.989757 | 70.181776 |

# Results

Once the work above is completed, we can compare some of statistics of the original regression analysis to the statistics derived using imputed data. Although the estimates obtained from the original analysis are close to the ones obtained using multiple imputation, replacing the missing data using multiple imputation gives us reason to believe we are getting a better model than if we had simply used the listwise deletion results.

Table - Comparing Original Results to Imputed Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter | Original Estimates | Original Std Err | Imputed Estimates | Imputed Std Err. | Original p-values | Imputed p-value |
| Cylinder | -1.19508 | 1.13851 | -0.998923 | 0.853992 | .3116 | .2423 |
| Size | 0.05221 | 0.02938 | 0.043529 | 0.024752 | .0973 | .0794 |
| Hp | -0.15009 | .07848 | -0.117772 | 0.053495 | .0765 | .0280 |
| Weight | -6.71776 | 3.98252 | -7.531299 | 3.641991 | .1138 | .0396 |
| Accel | -0.68451 | 3.98252 | -0.700246 | 0.411786 | .1423 | .0901 |
| Intercept | 67.61816 | 7.12819 | 66.361776 | 5.942839 | <.0001 | <.0001 |

The results of the work shown above give us a regression equation we can use to predict MPG based on the data we were presented as seen in Equation 1 below.

Equation - Final Regression Equation

It is difficult to find the overall p-value for the model, but with the greater degrees of freedom the statistical power is almost certainly improved in the new model compared to the initial fit with nearly half of the observations removed. A striking result from the imputation is how the statistical significance of every explanatory variable was improved. Many of the variables now seem to better account for the variation in MPG.

From these results it is our opinion that multiple imputation has a net benefit to improving the fit of the dataset. Now with the use of multiple imputation we find that the overall model is mostly definitely improved and the significance of each variable has improved.

# Future Work

While our results are promising it should be noted that multiple imputation should be used with consideration to the possible issues that may arise. While nearly half of the records were missing an observation, a single variable was not missing a majority of the observations. Something to note for future work is to pay special regard to variables that have a significant majority of the observations missing. Additional steps should be taken if the variable is necessary to the final model that is to be created.

There is also additional work in the subject of creating methods for implementing multiple imputation for binary and categorical variables. In a paper published through SAS there was demonstration of trying to impute binary values using the traditional linear imputation method, with noted success. [4] I was not able to find extensive coverage of the topic of imputing categorical data, so this currently seems like the biggest area for future work in the field of multiple imputation.

# References

|  |  |
| --- | --- |
| [1] | "SAS/STAT(R) 9.2 User's Guide, Second Edition: PROC Reg Missing Values," SAS, [Online]. Available: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug\_reg\_sect026.htm. [Accessed 7 September 2018]. |
| [2] | A. R. T. Donders, G. J. v. d. Heijden, T. Stijnen and K. G. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology,* vol. 59, no. 10, p. 1087–1091, 2006. |
| [3] | J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, M. G. Kenward, A. M. Wood and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ,* vol. 338, no. b2393, 2009. |
| [4] | Paul D. Allison, "Imputation of Categorical Variables with PROC MI," University of Pennsylvania, Philadelphia. |

# Appendix

SAS Code

/\* Michael Toolin \*/

/\* Steven Millett \*/

/\* Quantifying the world \*/

/\* Get the data \*/

FILENAME REFFILE 'C:\Users\Steven Millett\Dropbox\School\MSDS 7333 Quantifying the World\Session 2\carmpgdata\_26\_2.txt';

**proc** **import** datafile=REFFILE

out=mpg

dbms=dlm

replace;

delimiter='09'x;

**run**;

**PROC** **PRINT** data=mpg ;

**RUN**;

/\* Perform Regression on original data using list-wise deletion \*/

**PROC** **REG** DATA=mpg;

MODEL mpg = cylinders size hp weight accel;

**RUN**;

/\* Create table showing missing data \*/

ODS SELECT MISSPATTERN;

**PROC** **MI** DATA=mpg NIMPUTE=**0**;

VAR eng\_type cylinders size hp weight accel;

**run**;

/\* Create files with imputed data \*/

**PROC** **MI** DATA=mpg

OUT=MIOUT SEED=**1234**;

VAR eng\_type cylinders size hp weight accel;

**RUN**;

/\* Perform Regression on Imputed Data \*/

**PROC** **REG** DATA=MIOUT outest =outreg covout;

MODEL mpg = cylinders size hp weight accel;

by \_Imputation\_;

**RUN**;

**proc** **print** data=outreg;

**run**;

/\* Combine all Imputations \*/

**PROC** **MIANALYZE** DATA=outreg;

MODELEFFECTS cylinders size hp weight accel Intercept;

**RUN**;