Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

MSDS 7333 - Quantifying the World - Case Study 1 (Unit 2)

May 22, 2018

Abstract

It is said about 80% of a data scientist's job consists of cleaning data. This is because data sets often have either inaccurate or missing data that needs to be corrected or replaced respectively. The act of editing a data set affects the power, the probability that something is or isn't, and the bias, the likeliness of under or over estimating something. In this paper I am focusing on methods to complete data sets by imputing the missing values using regression techniques or deleting records with missing values. We will explain how the bias and power is affected by both and the different outcomes in the analysis they produce.

Introduction

In the data science community there is a saying garbage in garbage out. This describes how your inferences can only be as good as the data you are inferring about. Good data to analyze means the data set being analyzing has the correct information for the variables and a complete data set with no missing observations. In a perfect scenario, the data set you are analyzing will have complete correct observations for all variables. However, most of the time, data sets will have inaccurate or missing data.

The three patterns of missing data are "Missing at Random" (MAR), "Missing Completely at Random" (MCAR), and "Missing Not at Random". MAR is when there is missing data but it can be explained by other variables. An example of MAR would be someone answering a questionnaire and not answering a specific question do to other questions in that questionnaire. MCAR is like MAR with the exception that there is no correlation to why that observation is missing. The final pattern of missing data , MNAR, is when there is a pattern to the missing data or there is a reason why the data is missing.

When a data scientist needs to analyze a data set with missing data they can either impute the data or delete all rows with missing observations. When we decide to impute the data we fill in the missing observations based on the other observations we have. On the other hand, and a more simple form, when we delete a row we base this decision on if there is one or more missing observations then we remove the row completely.

In this case study, we are using a data set about cars that has 20 out of 38 records with missing observation for one or more variables. We will compare and contrast the methods of imputing and deleting observations and records respectively and the effects both methods have on the outcome.

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

Literature Review

Missing Data is a frequent phenomenon in the fields of Data Science and Data Analysis that can play a huge role in affecting the conclusions that can be drawn from our data set. There are many ways data can end up with missing values such as nonresponse, and dropout from the participant, or even due to mistakes from the researcher. As a result of that, identifying the pattern of missing data has become a very important part of data analysis. As noted in the work (1) of Trivellore Ranghunathan there are 3 patterns of missing data: Monotone, Univariate, and File Matching. In monotone patterns, a variable j is observed on a subset of subjects with variable j-1. This pattern is typically encountered in longitudinal or panel studies. In univariate patters, data is missing only on one variable in the analysis. This normally occurs in questionnaires. In file matching patterns, we have a collection of 2 files that are appended, causing a casual inference between 2 variables Y1 and Y2. This pattern can be typically encountered in treatments.

In Steff van Buuren book (2) the concepts of incomplete data are discussed. Missing data can be clustered into 3 categories: Missing completely at random (MCAR), Missing at random (MAR), and Missing not at random (MNAR). Van Buuren's notes that, if the missingness of data occurs entirely at random, independent of observed and unobserved parameters, then our data are missing completely at random (MCAR). However, data are rarely MCAR. When the missing data is not random but can be accounted for by variables that exist in the data set, then our data are missing at random (MAR). Finally, if the data is not part of the other 2 categories, then it can be considered missing not at random (MNAR).

Background

This dataset holds information regarding automobiles. Our goal is to compare the effects of listwise deletion and multiple imputation on a regression analysis. The response variable will be miles per gallon (MPG), and the explanatory variables are auto, cylinders, size, hp, weight, accel, and eng_type. This set is comprised of 8 elements and 38 observations. From our total data, only 18 records can be used in our analysis using listwise deletion since they have no missing values. The rest contain one or more missing variables and thus some sort of imputation is required. Detailed information of all variables follows below:

Table 1: Variable Descriptions

Variable	**Regression**	**Description**	**Type**
MPG	Response	Estimated miles per gallon	Continuous
Auto	Explanatory	Make and model of car	String
Cylinders	Explanatory	Number of engine cylinders	Discrete

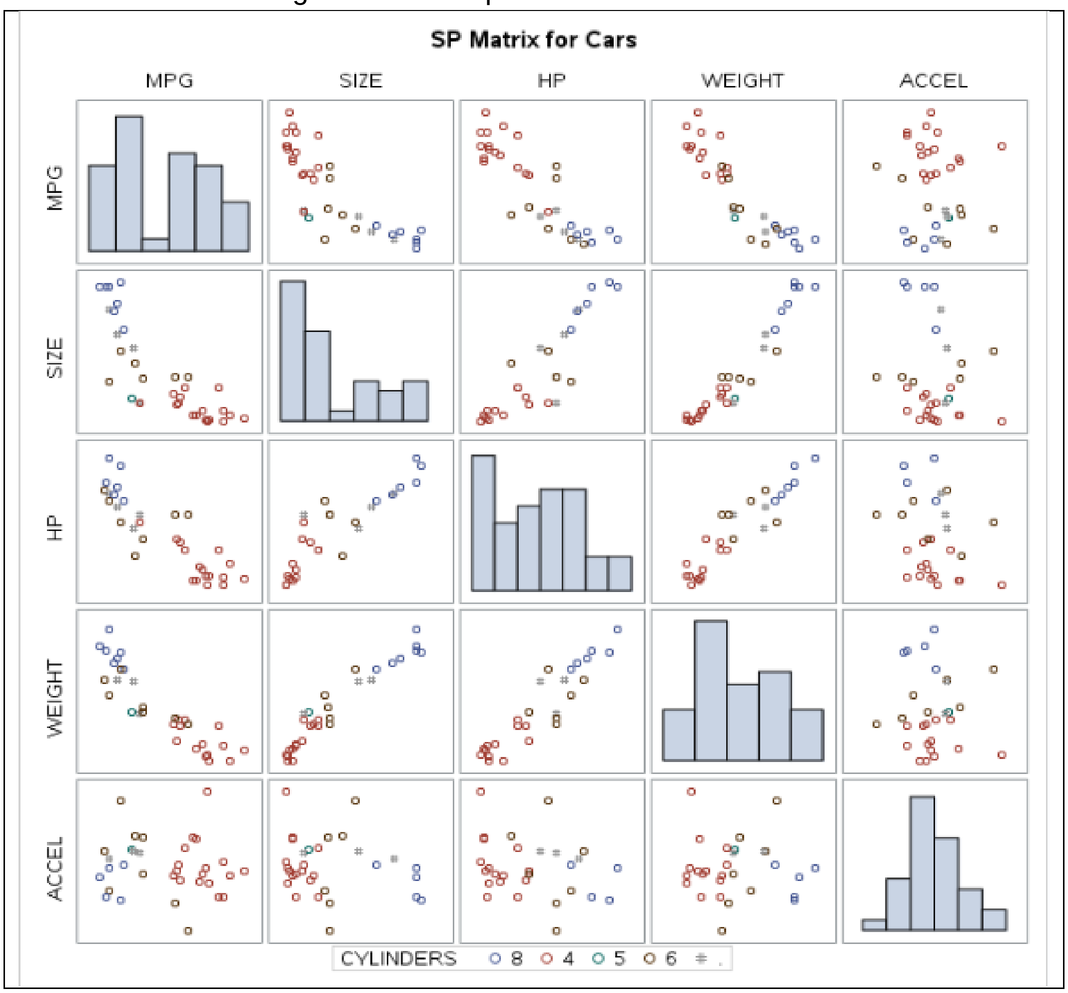
Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

Size	Explanatory	Engine displacement measurement	Continuous
HP	Explanatory	Engine horsepower	Continuous
Weight	Explanatory	Weight of vehicle	Continuous
Accel	Explanatory	Acceleration	Continuous
Eng_Type	Explanatory	Engine type	Discrete

Visual inspection of the raw data depicts powerful linear correlations. That can be reaffirmed by the correlations matrix. Linear relationships between weight – mpg, size – mpg and hp – mpg can be easily deduced from the scatterplot matrix grouped by Cylinders as seen below:

Figure 2: Scatterplot Matrix for Autos



Almost all our explanatory variables depict normal distributions. The only exceptions are miles per gallon (mpg) and size; however, no transformations were performed. The variable auto can be compared to a unique id type of variable that ends up being irrelevant in our analysis, hence we came to the decision of removing it. For the variables size and cylinders a case can be made of notable correlation (0.95) between them. Nevertheless, we decided to retain all variables and continue with our analysis.

Methods

First we used likewise deletion to remove any record having at least one missing observation. We accomplished this in SAS by using a multiple linear regression (MLR) which, by default, automatically removes any record having a missing observation.

In our MLR we used the variables “CYLINDERS”, “SIZE”, “HP”, “WEIGHT”, “ACCEL”, and “ENG_TYPE” to predict MPG. As we can see, of the 38 observation, the model only used 18 and deleted 20. The impact of deleting over 50% of our records is the power of our model is dramatically less, with 17 degrees of freedom, than when we imputed the data which we will see later on.

Table 3: Observations Read, Used, and Deleted.

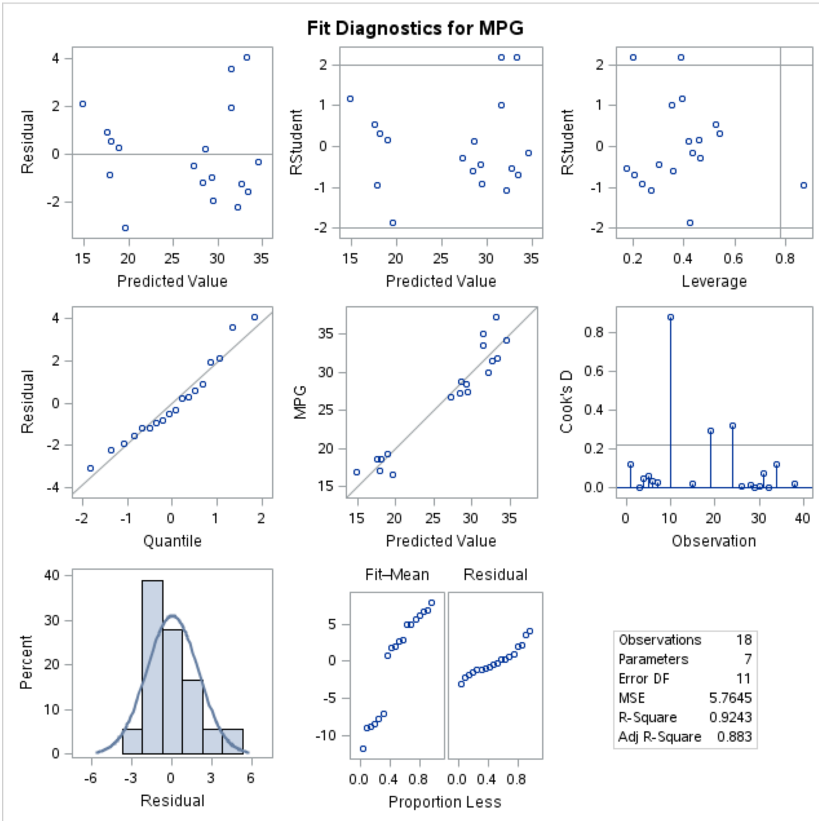
Number of Observations Read	38
Number of Observations Used	18
Number of Observations with Missing Values	20

Furthermore, we will not need to transform our data as we can see the normality in the histogram and residual plots tin the Fit Diagnostic for MPG table below.

Table 4: Fit Diagnostic for MPG

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis



Next we imputed the data set with proc MI. Again, we used “CYLINDERS”, “SIZE”, “HP”, “WEIGHT”, “ACCEL”, and “ENG_TYPE” but we also used “MPG” as our variables. Unlike listwise deletion, which simply removes all records with a missing observation, when imputing data, you first need to understand how the data is missing. As we can see in the table below “Finding Missing Patter” , there is not a pattern to the missing data and the data that is missing cannot be explained by other variables meaning there is no correlation with a missing data point with another data point. From these inferences, we can conclude are data is MCAR, or missing completely at random.

Table 5: Finding Missing Pattern

Group	MPG	Cyl	Size	HP	Weight	Accel	Eng_Type	Freq	Pct
1	X	X	X	X	X	X	X	18	43.37
2	X	X	X	X	X	X		2	5.26
3	X	X	X	X	X		X	1	2.63
4	X	X	X	X	X			1	2.63
5	X	X	X	X		X	X	3	7.89
6	X	X	X	X			X	1	2.63
7	X	X	X		X	X	X	5	13.16

8	X	X		X	X	X	X	2	5.26
9	X	X		X		X	X	1	2.63
10	X		X	X	X	X	X	2	5.26
11	X		X	X	X		X	1	2.63
12	X		X	X		X	X	1	2.63

Table 6: Finding Missing Pattern –Means

GROUP	MPG	Cyl	Size	HP	Weight	Accel	Eng_Type
1	26.61	5.33	177.06	101.89	2.8	14.36	.33
2	31.35	4	95	70	2.13	16.85	
3	18.20	8	318	135	3.8		1
4	17.60	8	302	129	3.73		
5	28.13	4.67	128	72.67		16.17	0
6	21.50	4	121	110			0
7	22.32	5.4	182.8		3	15.24	.4
8	19.10	6		115	3.11	15.15	0
9	30.50	4		78		14.1	0
10	21.10		176	110	3.08	15.75	0
11	18.10		258	120	3.41		0
12	17.00		305	130		15.4	1

For data sets with monotone missing data patterns, data that can be based off of previous observations, you can use monotone methods to impute missing values for the variables. For continuous variables like HP and WEIGHT, regression methods are better suited, whereas for categorical variables like CYLINDERS and ENG_TYPE, discriminant function methods are better suited to classify the variable. (1).

Now that we know how the data is missing and the method to use to impute the data, we use MCMC to impute the data. We set the seed at 100 with 5 imputations. This gives us 5 data sets with degrees of freedom of 37, which shows we these model have more power then are listwise deletion method which only has 17 degrees of freedom. This means we can have more confidence that are model using the imputed data will give us a stronger probability. Now that we have our complete data sets we can now compare our models using our listwise deletion data and our multiple imputation data set and the pros and cons of both.

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

Results

Looking at the mean values of all our variables we do not discern any major differences between likewise deletion and multiple imputation. However, comparing the standard errors we do see some big variations. As expected multiple imputation lowers our standard error across board, especially on variables size and hp. That comes as no surprise since adding supplementary observations based on the data present, will reduce the variance of our parameters. Detailed information can be seen on the table below:

Table 7: Parameter Comparison for Imputation Methods

	Listwise Deletion		**Multiple Imputation**	
Variables	Mean	Std Error	Mean	Std Error
CYLINDERS	5.32	0.28	5.39	0.11
SIZE	180.89	15.45	180.45	6.38
HP	101.33	4.72	101.98	1.89
WEIGHT	2.91	0.13	2.86	0.05
ACCEL	14.94	0.27	14.93	0.11
ENG_TYPE	0.29	0.08	0.27	0.03

NOTE: We need to be careful of the function PROC MI. This function normally generates continuous values, and that can cause an issue if we are imputing on a discrete variable. This can be seen on the missing cylinders data for Saab 99 GLE, Buick Century Spec, AMC Concord D/L, and Chevy Caprice Classic. Details can be seen on table below:

Table 8: MCMC Results for Discrete Variable Cylinders

Imputation	**Auto**	**cylinders**
1	Saab 99 GLE	4.74
1	Buick Century Spec	4.84
1	AMC Concord D/L	6.53

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

1	Chevy Caprice Classic	7.1

As we can see from the table above, the imputed data for cylinders does not make any sense. For all intents and purposes, cylinders are supposed to be integers with a range of 4-12. This is normal when using the MCMC process since it expects continuous variables. A solution to this would be to round up or down to the nearest integer. Nonetheless, linear rounding is known to perform badly during mean estimation, and supplementary logistic methods were not applicable to this dataset, so a decision was made to keep the continuous variables.

For our regression analysis, our response variable will be MPG. Our explanatory variables will be CYLINDERS, SIZE, HP, WEIGHT, ACCEL, and ENG_TYPE. We will perform multiple linear regression using both listwise deletion and multiple imputation. Fitting the model using listwise deletion yields a set with a total of 18 variables, whereas fitting on an imputed data set yields 38 which is preferable due to the small size of this data set. More information regarding listwise deletion procedure can be seen on the table below:

Table 9: Multiple Linear Regression with Listwise Deletion

Analysis of Variance (listwise deletion)					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	774.2	129	22.4	<0.0001
Error	11	63.4	5.8		
Corrected Total	17	837.7			

R-Square	0.924
Adj R-Sqr	0.883

Parameter Estimates (listwise deletion)			
Variable	**Parameter Estimate**	**Standard Error**	**PR >|t|**
Intercept	70.14	8.01	<0.0001
CYLINDERS	-3.33	1.56	0.056
SIZE	0.02	0.03	0.49

HP	-0.2	0.08	0.03
WEIGHT	-0.31	5.13	0.95
ACCEL	-0.78	0.58	0.21
ENG_TYPE	6.6	3.59	0.09

Despite only having 18 usable observations, the listwise deletion procedure yields a model with an F-Value = 22.4 an R-Square = 0.924 and an Adjusted R-Square = 0.883. Despite the small number of observations, this model explains a notable amount of variance. An issue that arises though is collinearity on variables such as weight and size, which does not surprise, and that's the reason we wish to examine the results of listwise imputation. Output of the listwise imputation procedure can be seen on table below:

Table 10: Multiple Linear Regression with Listwise Imputation (5 Models)

Parameters	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Pr > F	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
F Value	51.8	42.3	35.9	52.3	37
R-Square	0.909	0.891	0.874	0.91	0.877
Adj R-Square	0.891	0.87	0.85	0.892	0.854

Five imputed data sets were created using command PROC MI with MCMC. For each imputed dataset regression analysis was performed. Information for each variable (cylinders, size, horsepower, weight, acceleration, and engine type) of each model are shown below.

Table 11: Regression Coefficients for Multiple Imputation (5 Models)

	Model 1			**Model 2**		
Variable	**Parameter Estimate**	**Standard Error**	**PR >|t|**	**Parameter Estimate**	**Standard Error**	**PR >|t|**
Intercept	67.3	4.7	< 0.0001	66.1	4.9	< 0.0001
Cylinders	-1.8	0.8	0.03	-2.3	0.9	0.02
Size	0.5	0.01	0.02	0.04	0.02	0.05
HP	-0.1	0.05	0.03	-0.11	0.05	0.02
Weight	-8.1	2.8	0.01	-7.3	3.1	0.03
Accel	-0.5	0.3	0.11	-0.4	0.3	0.2
Eng Type	3	1.6	0.07	3.9	1.6	0.02

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

	Model 3			Model 4		
Variable	**Parameter Estimate**	**Standard Error**	**PR >|t|**	**Parameter Estimate**	**Standard Error**	**PR >|t|**
Intercept	67.1	6.1	< 0.0001	65.7	4.3	< 0.0001
Cylinders	-1.9	0.97	0.05	-2.5	0.8	0.01
Size	0.05	0.02	0.03	0.04	0.02	0.04
HP	-0.15	0.05	0.01	-0.012	0.05	0.02
Weight	-6.5	3.1	0.05	-6.5	3.1	0.04
Accel	-0.5	0.4	0.2	-0.34	0.3	0.2
Eng Type	2.8	1.7	0.1	4.7	1.6	0.01

	Model 5		
Variable	**Parameter Estimate**	**Standard Error**	**PR >|t|**
Intercept	67.1	5.3	< 0.0001
Cylinders	-1.6	0.9	0.09
Size	0.02	0.02	0.3
HP	-0.2	0.05	0.01
Weight	-3.7	2.8	0.2
Accel	-0.8	0.37	0.04
Eng Type	3.3	1.8	0.07

As we can see from the tables above, there is a significant difference for variables weight and acceleration. In the multiple imputed regression model, variable weight becomes almost significant (p=0.5). The most heavily imputed variables (size, weight, cylinders) show an improvement in standard error and a decrease in variance as is expected of an imputed model. Additionally, the imputed regression model performs slightly better than the listwise deletion model. Despite the extra observations, the two models performed about the same with a root mean square error of MSE = 5.9 (imputed) and MSE = 5.8 (deletion) and Adj R-Square values of Adj R-Square = 0.87 (imputed) and Adj R-Square = 0.88 (deletion).

Table 12: Comparison of Important Statistics

Parameters	**Mean of imputed Models**	**Listwise deletion**
Pr > F	<0.0001	<0.0001

F Value	43.86	22.4
R-Square	0.8922	0.924
Adj R-Sqr	0.8714	0.883
MSE	5.9	5.8

Future Work, Discussion Conclusions, and Next Steps

There are different scenarios to use listwise deletion as well as multiple imputation for data sets. Listwise deletion may be a good method when dealing with big data, or data that needs non-traditional methods of processing and analysis. As an example, if this cars data set had a billion records, and we deleted 50%, we would still have 500 million records to analyze. In this case, we would have enough records to draw strong inferences from an analysis. On the other hand, if the data set we are working with is relatively small, like the one we are using, than multiple imputation may be a good solution to complete the data set for analyzation.

In this vein, imputation may not be needed as much in the future for completing data sets. This is because as the volume and velocity of data increases, and with the invention of new tools to aggregate these data, data scientist may have more data sets with enough complete records to analyze without the need of imputation. However, with new machine learning algorithms in regards to facial recognition and language processing, the concepts of imputing missing data can be applicable to make the models more robust with predictions.

In our case study, the cars data set was a good scenario to use the imputation method. The imputation method was beneficial because it gave our model more power, with more than double the degrees of freedom. Even though the listwise deletion data marginally explained the variance better, with a lower mean square error and higher adjusted R-square, we can be more confident with the probability of the predictions using the imputation data set because of the extra degrees of freedom.

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

References:

1. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect020.htm
2. Trivellore Raghunathan. “Missing Data Analysis in Practice.” CRC Press, 2016.
3. Steff van Buuren. “Flexible Imputation of Missing Data.” Chapman & Hall/CRC, 2012

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

Appendix - SAS Code

```
data cars;
infile '/home/atschannen0/sasuser.v94/carmpgdata_26 (2).txt' dlm= '	'
firstobs=2 dsd;
length Auto $30;
input Auto MPG    CYLINDERS   SIZE  HP    WEIGHT      ACCEL ENG_TYPE ;
proc print data=cars;
run;


*Use PROC MI to discover the missing values
patterns and to decide what MI options to use;

Title 'Identifying Missing Patterns with PROC MI';
ODS SELECT MISSPATTERN;
PROC MI data = cars seed = 100 nimpute = 0;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
MCMC;
RUN;
QUIT;

PROC MI DATA = cars OUT = carout SEED = 100 NIMPUTE = 5;
VAR CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
MCMC; *imputes for arbitrary missing pattern;
RUN;
QUIT;

PROC PRINT data = carout;
RUN;

PROC SORT data = carout;
BY _imputation_;
RUN;



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

PROC REG data = cars;
MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
OUTPUT OUT = regcars
RSTUDENT = rstudent
COOKD = cookd;
RUN;
QUIT;
```

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis

```
TITLE 'MLR on Each Imputed Dataset';
PROC REG DATA = carout OUTEST = outreg COVOUT;
MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
BY _IMPUTATION_;
RUN;
QUIT;


PROC MIANALYZE DATA = outreg;
MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE Intercept;
RUN;


* likewise deletion;
Title 'Predicting MPG with Listwise Deletion Imputation';
PROC REG data = cars;
MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
OUTPUT OUT = carsdelet
RSTUDENT = rstudent
COOKD = cookd;
RUN;
QUIT;


*exploring averages ;
*aveages for likewise;
PROC MEANS data = carsdelet N NMISS MEAN STD STDERR CLM Q1 MEDIAN Q3;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;

*averages for mlr;

PROC MEANS data = regcars N NMISS MEAN STD STDERR CLM Q1 MEDIAN Q3;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL;
RUN;



*scatter plot class by cylinders ;
PROC SGSCATTER data = cars;
title 'SP Matrix for Cars';
matrix MPG  SIZE HP WEIGHT ACCEL / group=CYLINDERS diagonal=(histogram);
RUN;
```

Multiple Imputation of Missing Data - Car Fuel Efficiency Data

Evangelos Giakoumakis