Abstract: I conducted a comparative study to compare results between listwise deletion and multiple imputations. The original dataset had 38 observations of which 20 were missing. I used SAS PROC MI procedure to determine the pattern of missingness and it turned out to be arbitrary (aka non-monotone). I decided to use MCMC (Markov Chain Monte Carlo) as the imputation method as it is recommended for continuous and arbitrary missing patterns[1]. The number of imputations was set at 5. Linear regression model was fitted for each imputed dataset. These regression models used all 38 observations. SAS procedure PROC MIANALYZE was used to combine results of these 5 regression models, and then I compared the combined model to the original model. Only HP was included in the regression model with original data while regression model with imputed data included HP, CYLINDERS, and ENG_TYPE variables.

Introduction: This case study will compare linear regression models fitted on the datasets after listwise deletion and multiple imputations. In listwise deletion, the entire record is omitted if there is missing value in any of the columns of the records. This results in a reduction in statistical power and smaller sample size. On the other hand, in multiple imputations, the missing values are imputed by averaging multiple imputes which are generated based on characteristics of the original datasets.

Literature Review: There are multiple imputations methods available such as Listwise (aka complete case) deletion, pair wise deletion, hot deck imputation, cold deck imputation, mean substitution, weighting methods, multiple imputations, etc. Out of these imputations, multiple imputation fairs better than others because it doesn't drop record from the model and, it doesn't alter the characteristics of the original data and, it maintains the uncertainty in the missing values in the final model. Multiple imputations follow three steps 1. Imputation 2. Analysis 3. Pooling.[2] It can be performed when data is missing completely, missing completely at random, and missing not at random.

Method: The steps used for this analysis were: (1) linear regression on the original dataset with listwise deletion, (2) the missing data pattern was analyzed (3) multiple imputations was performed to impute the missing values (4) individual regression was ran on each imputed data, (5) the multiple imputation results were combined and averaged, and (6) the models of listwise deletion and multiple imputation were compared.

Results:

(1) **Linear regression on the original dataset with listwise deletion.** The SAS code used for running linear regression was:

```
PROC REG DATA = WORK.IMPORT1;
        MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;
```

[1] SAS Imputation Method Recommendations

[2] Multiple Imputation Wiki

I observed that more than half of observations were dropped from the model. 20 observations out of 38 were dropped from the model. Only 18 observations were used in the model. This drastically reduces the statistical power and accuracy of the model. This regression model obtained here will be compared to the imputed regressions model generated.

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

(2) **Analyze the Missing Data Pattern:** The SAS code snippet for analyzing the missing data pattern is:

```
ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.IMPORT1 NIMPUTE=0;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;
```

The missing data pattern was found to be arbitrary, or non-monotone.

Missing Data Patterns									
Group	**MPG**	**CYLINDERS**	**SIZE**	**HP**	**WEIGHT**	**ACCEL**	**ENG_TYPE**	**Freq**	**Percent**
1	X	X	X	X	X	X	X	18	47.37
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

As per arbitrary pattern of missingness, the recommend imputation method was MCMC full-data imputation.

(3) **Perform Multiple Imputations to Impute the Missing Values:** The SAS code snipped used for performing the multiple imputation is shown below. MCMC is the default method for SAS Proc MI. I

have set the number of imputation at 5. The output of imputations is stored in dataset WORK.IMPORT1. This dataset will have 5 different imputed datasets.

```
PROC MI DATA = WORK.IMPORT1
OUT = MIOUT seed = 35399 NIMPUTE=5;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;
```

Below SAS output table shows the results of multiple imputations

Model Information	
Data Set	WORK.IMPORT1
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	35399

(4) **Regression with Each Imputed Data:** The SAS code snipped used for performing regression on each imputed dataset:

```
PROC REG data = miout outest = outreg covout ;
Model MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
by _Imputation_;
RUN;
```

in the above code, “by” statement asks SAS to do regression on each imputed dataset. Also, it uses all the available datasets i.e. all 38 records as shown in the below table.

Number of Observations Read	38
Number of Observations Used	38

The parameter estimates for each of the listwise model and five models are shown in the below table

Variable	Original	1	2	3	4	5
Intercept	70.14772	71.07423	72.39109	68.6843	68.5524	67.01166
CYLINDERS	-3.33403	-3.03737	-2.9346	-2.93177	-2.85923	-2.69887
SIZE	0.0228	0.02391	0.02052	0.02557	0.04358	0.04106
HP	-0.19546	-0.15919	-0.19765	-0.14873	-0.15208	-0.13697
WEIGHT	-0.30623	-2.03889	-0.26845	-2.97682	-4.65964	-6.12984
ACCEL	-0.78199	-0.91547	-1.07191	-0.69925	-0.57066	-0.35255
ENG_TYPE	6.5988	5.74751	6.22872	5.80842	5.19245	6.29941

(5) **Multiple Imputation Results Combined:** The SAS command for combing the outputs of the regression analysis is PROC MIANALYZE. The code used to combine the 5 different imputations is shown below.

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

The table below shows that parameter estimates, standard errors and p-value for hypotheses test for each explanatory variable in the original regression model and imputed regression model. The original regression was based on 18 observations whereas the multiple imputations based on all 38 observations. Based upon a cut-off of 0.05 for Pr > |t| only HP was included in the regression model with original data (listwise deletion) while regression model with imputed data included HP, CYLINDERS, and ENG_TYPE.

	Parameter Estimates		Standard Errors		Pr > |t|	
Variable	Original	Imputed	Original	Imputed	Original	Imputed
Intercept	70.14772	69.542738	8.03838	4.676262	<.0001	<.0001
CYLINDERS	-3.33403	-2.892369	1.56072	0.766712	0.056	0.0002
SIZE	0.0228	0.030931	0.03207	0.021663	0.4918	0.1597
HP	-0.19546	-0.158924	0.08065	0.046085	0.0338	0.0013
WEIGHT	-0.30623	-3.214728	5.13263	3.740139	0.9535	0.4001
ACCEL	-0.78199	-0.721966	0.58264	0.409842	0.2066	0.103
ENG_TYPE	6.5988	5.855301	3.59008	1.579569	0.0932	0.0002

Discussion & Future Work: With SAS MI procedure, I could impute missing data and with SAS MIANALAZE procedures I combined and averaged estimates for regression coefficients for linear regression based on 5 different imputes. Next steps could be to replicate the case study using R and Python languages.

Appendix: SAS Code

```
FILENAME REFFILE "/folders/myfolders/Data/carmpgdata.txt";

PROC IMPORT DATAFILE=REFFILE
dbms=tab
out= WORK.IMPORT1;
getnames= yes;
run;

PROC CORR DATA=WORK.IMPORT1 PLOTS=MATRIX(HISTOGRAM NVAR=7);
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;

PROC REG DATA = WORK.IMPORT1;
        MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;

ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.IMPORT1 NIMPUTE=0;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;

PROC MI DATA = WORK.IMPORT1
OUT = MIOUT seed = 35399 NIMPUTE=5;
VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN;

PROC REG data = miout outest = outreg covout ;
Model MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
by _Imputation_;
RUN;

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