Car MPG Case Study

Amit Patel, Sanjay Jagarlamudi, Murali Parthasarathy
MSDS 7333 Quantifying the World, May-20-2017

ABSTRACT

Environment and consumer awareness are driving car OEMs to shift focus on increasing the miles per gallon offered by the vehicle. There are several factors, which affect the MPG of a vehicle. The data collected is from the various sub systems of the car and have a lot of missing values. The missing values are caused by the devices, which transfer the data from the car into a human consumable format. The Paper studies the various options available for data imputation to increase the effectiveness of the MPG data analysis.

INTRODUCTION

The vehicle mpg is highly dependent on the built of the vehicle and the driving style. Vehicles mpg is drastically reduced in city drives where the driver has to slower speed limits and multiple signal lights. The driver has to accelerate to gain momentum after every slow down or signal light. The MPGs in city drives are affected by driving styles and vehicle attributes. The data set provided has a lot of missing values, which affect the overall statistical capability of the data set. The objective of the study is to find affective ways of data imputation.

Data Analysis

Attribute Name	**Attribute Type**	**Characteristics**
Mpg	Continuous	Miles Per gallon. This is the reading from the car during the drive
Cylinders	Categorical - Ordinal	The number of cylinders in the car determines the mpg
Hp	Continuous	Horse Power. The horse power generated by the engine
Weight	Continuous	Weight of the car. A heavier car will need more displacement to gain momentum

Acceleration	Continuous	Acceleration is a driving style. It depends on the pressure exerted by the driver on the accelerator gear
Eng_type	Categorical - Ordinal	The Engine type, the type of engine could be a V4 or V6

Table 1: Data attributes

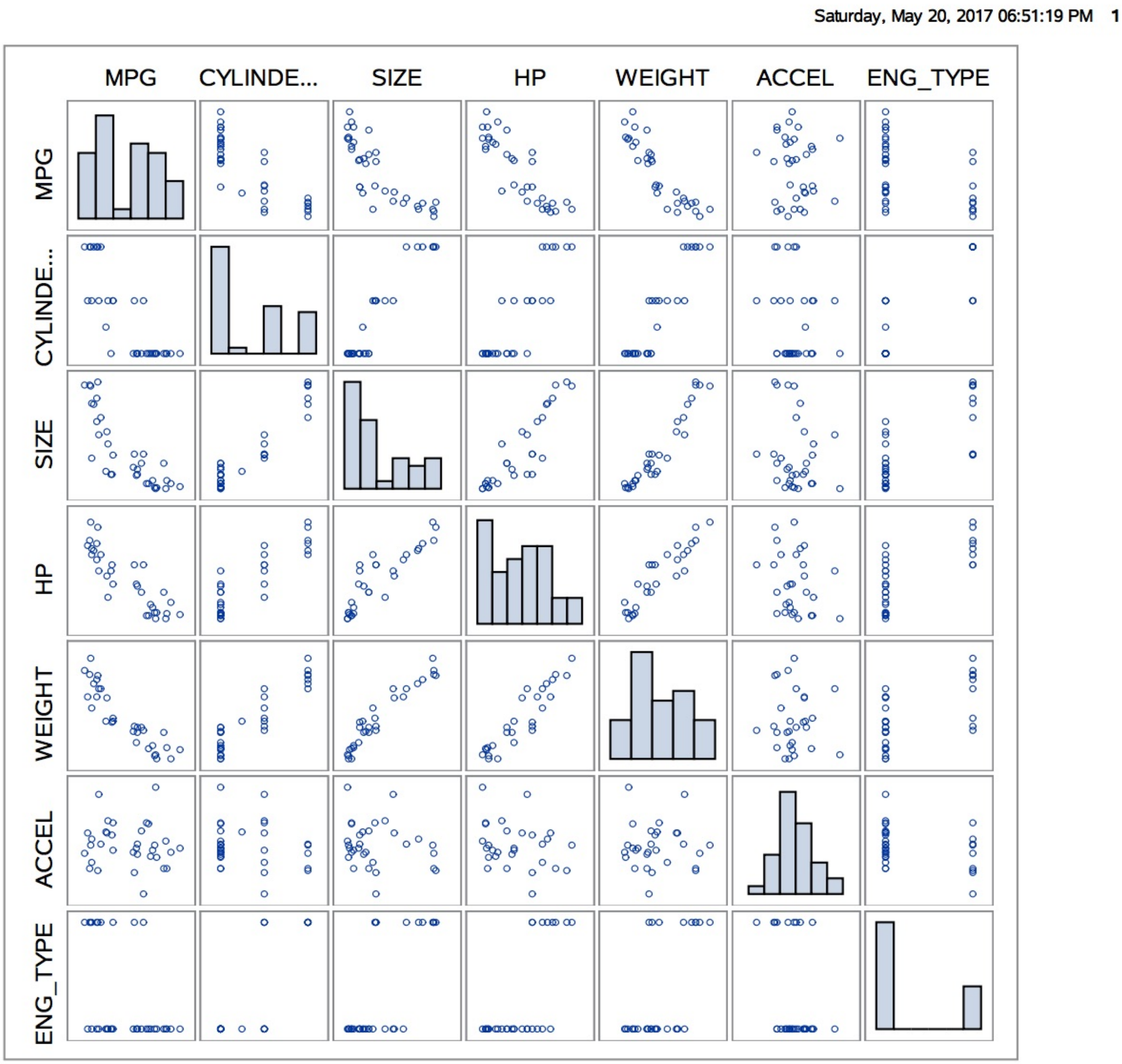

Figure 1 : Scatter Plot

The section below comments a few relationships

Relationship	**Comments**
Mpg-weight	The mpg decreases as the weight of the vehicle increases.
Hp-mpg	The mpg decreases as the vehicle offers higher Horsepower
Weight-hp	The horse power increases as the weight of the vehicle increases

Table 2: Data correlation

Missing Values

The data set has a lot of missing values. A cursory inspection suggests that weight, HP and engine type are missing on some data records. The missing values display a monotonous pattern; the figure below provides insight to the missing value monotonous pattern.

The MI Procedure

Missing Data Patterns

										Group Means						
Group	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE	Freq	Percent	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE
1	X	X	X	X	X	X	X	18	46.15	26.605556	5.333333	177.055556	101.888889	2.795333	14.355556	0.333333
2	X	X	X	X	X	X	.	2	5.13	31.350000	4.000000	95.000000	70.000000	2.125000	16.850000	.
3	X	X	X	X	X	.	X	1	2.56	18.200000	8.000000	318.000000	135.000000	3.830000	.	1.000000
4	X	X	X	X	X	.	.	1	2.56	17.600000	8.000000	302.000000	129.000000	3.725000	.	.
5	X	X	X	X	.	X	X	3	7.69	28.133333	4.666667	128.000000	72.666667	.	16.166667	0
6	X	X	X	X	.	.	X	1	2.56	21.500000	4.000000	121.000000	110.000000	.	.	0
7	X	X	X	.	X	X	X	5	12.82	22.320000	5.400000	182.800000	.	3.009800	15.240000	0.400000
8	X	X	.	X	X	X	X	2	5.13	19.100000	6.000000	.	115.000000	3.112500	15.150000	0
9	X	X	.	X	.	X	X	1	2.56	30.500000	4.000000	.	78.000000	.	14.100000	0
10	X	.	X	X	X	X	X	2	5.13	21.100000	.	176.000000	110.000000	3.087500	15.750000	0
11	X	.	X	X	X	.	X	1	2.56	18.100000	.	258.000000	120.000000	3.410000	.	0
12	X	.	X	X	.	X	X	1	2.56	17.000000	.	305.000000	130.000000	.	15.400000	1.000000
13	O	O	O	O	O	O	O	1	2.56	.	.	.	.	.	.	.

Figure 2: Missing attribute Pattern

Missing data pattern

There are four data missing mechanisms
Missingness completely at random - The probability of missingness is the same for all variables
Missingness at random - The probability of missingness is not the same for all variables

Missingness depends on observed predictors –The missingness is not at random and depends on the observed values

The data set shows a patterns of missing at random, the probability of missingness is not the same for all the variables.

Pattern of Missingness	Type of Imputed Value	Recommended Algorithm
Missingness at random (MAR)	Continuous	MCMC
	Categorical	MCMC

Also through observation we are assuming the data is multivariate normal distribution.

Imputation Methods

The imputation methods can be categorized as

- Simple imputation
- Multiple imputations

The methods can be further categorized as approaches, which keep all the data, and approaches will remove records with missing data.

Simple imputation

Complete case Analysis

In the method the records with missing data is completely ignored. This method will have bias as certain good responses with missing values will be ignored

The other options available are

Last Value Carry Forward – Use the last observed value. This approach will not be suitable for the car data set as the attributes are highly driven by the driving style. This method will introduce a lot of bias to the data set

Hot Deck imputation – This method uses the nearest neighbor concept where the closest value is substituted. Again this approach may not work as the missing attributes are dependent on the driving style.

Single Imputation Methods

The dataset is highly co related but there are certain attributes, which will qualify for simple imputation methods and will not affect the overall predictability of the model.

Single imputation approaches can be done for engine type and weight. The most likely value can be determined from the data available. The approach is not comprehensive and cannot be employed to all the missing attributes.

Multiple Imputation Approach

The multiple imputation technique determines the missing values by taking into account the available data sets have to be employed. The multiple imputation technique takes into account the uncertainty in the data set due to the missing values. The multiple imputation technique is easy to employ.

The only disadvantage of the imputation model is the team now has to consider both the analysis model and imputation model.

Multiple data sets approach

There are two approaches they are

- Maximum likelihood estimation
- Multiple imputations

Maximum likelihood estimation

The maximum likelihood estimation does not create multiple data sets like the multiple imputation method. The method estimates the missing value based on the available attributes. This approach fits linear models very well. The approach estimates the values based on a range of values.

Regression Analysis (List wise/Complete)

The regression analysis on the data set shows only 17 out of the 39 data records are used. **The p-values of the coefficients are greater than 0.05, which reduces the power of our model to estimate Mpg.** To increase the power of the model, we will add observations with missing values to see if it improves our estimates.

The REG Procedure
Model: MODEL1
Dependent Variable: MPG

Number of Observations Read	39
Number of Observations Used	18
Number of Observations with Missing Values	21

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	774.27999	129.04667	22.39	<.0001
Error	11	63.40945	5.76450		
Corrected Total	17	837.68944			

Root MSE	2.40094	R-Square	0.9243
Dependent Mean	26.60556	Adj R-Sq	0.8830
Coeff Var	9.02419		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.14772	8.03838	8.73	<.0001
CYLINDERS	1	-3.33403	1.56072	-2.14	0.0560
SIZE	1	0.02280	0.03207	0.71	0.4918
HP	1	-0.19546	0.08065	-2.42	0.0338
WEIGHT	1	-0.30623	5.13263	-0.06	0.9535
ACCEL	1	-0.78199	0.58264	-1.34	0.2066
ENG_TYPE	1	6.59880	3.59008	1.84	0.0932

Figure 3 Linear Regression

Multiple Imputation Details

Step 1 – Create the Data Sets

The MI Procedure

Model Information	
Data Set	WORK.CASESTUDY1
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

Figure 4 Multiple Imputation Procedures

Variance Information (5 Imputations)							
	Variance						
Variable	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
CYLINDERS	0.000238	0.067196	0.067481	34.996	0.004242	0.004233	0.999154
SIZE	0.437624	209.591060	210.116208	35.06	0.002506	0.002502	0.999500
HP	0.144206	18.784569	18.957616	34.804	0.009212	0.009169	0.998169
WEIGHT	0.000063479	0.012967	0.013043	34.934	0.005874	0.005857	0.998830
ACCEL	0.003511	0.065708	0.069921	32.07	0.064126	0.061963	0.987759
ENG_TYPE	0.000081952	0.005427	0.005525	34.43	0.018122	0.017955	0.996422

Parameter Estimates (5 Imputations)										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
CYLINDERS	5.408733	0.259770	4.8814	5.9361	34.996	5.395728	5.431822	0	20.82	<.0001
SIZE	179.554826	14.495386	150.1294	208.9802	35.06	179.069351	180.695923	0	12.39	<.0001
HP	103.056213	4.354035	94.2153	111.8972	34.804	102.544009	103.555236	0	23.67	<.0001
WEIGHT	2.864336	0.114207	2.6325	3.0962	34.934	2.850770	2.871197	0	25.08	<.0001
ACCEL	14.901253	0.264426	14.3627	15.4398	32.07	14.810038	14.952625	0	56.35	<.0001
ENG_TYPE	0.285951	0.074331	0.1350	0.4369	34.43	0.275788	0.296527	0	3.85	0.0005

Figure 5 Parameter and Variance Estimation on Imputed value

Step 2 - Analyze the imputed data set

In this step a regression analysis is done on the imputed data sets.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1475.92029	245.98672	69.22	<.0001
Error	31	110.17050	3.55389		
Corrected Total	37	1586.09079			

Root MSE	1.88518	R-Square	0.9305
Dependent Mean	24.76053	Adj R-Sq	0.9171
Coeff Var	7.61363		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.07423	4.09818	17.34	<.0001
CYLINDERS	1	-3.03737	0.75954	-4.00	0.0004
SIZE	1	0.02391	0.01830	1.31	0.2010
HP	1	-0.15919	0.03985	-3.99	0.0004
WEIGHT	1	-2.03889	2.81135	-0.73	0.4737
ACCEL	1	-0.91547	0.27746	-3.30	0.0024
ENG_TYPE	1	5.74751	1.43032	4.02	0.0003

Figure 6 Imputation 1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	1480.23596	246.70599	72.25	<.0001
Error	31	105.85483	3.41467		
Corrected Total	37	1586.09079			

Root MSE	1.84788	**R-Square**	0.9333
Dependent Mean	24.76053	**Adj R-Sq**	0.9203
Coeff Var	7.46302		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	72.39109	4.09768	17.67	<.0001
CYLINDERS	1	-2.93460	0.70873	-4.14	0.0002
SIZE	1	0.02052	0.01918	1.07	0.2930
HP	1	-0.19765	0.04055	-4.87	<.0001
WEIGHT	1	-0.26845	3.07210	-0.09	0.9309
ACCEL	1	-1.07191	0.28020	-3.83	0.0006
ENG_TYPE	1	6.22872	1.37736	4.52	<.0001

Figure 7 Imputation 2

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	1478.18064	246.36344	70.77	<.0001
Error	31	107.91015	3.48097		
Corrected Total	37	1586.09079			

Root MSE	1.86574	**R-Square**	0.9320
Dependent Mean	24.76053	**Adj R-Sq**	0.9188
Coeff Var	7.53512		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	68.68430	3.61127	19.02	<.0001
CYLINDERS	1	-2.93177	0.72111	-4.07	0.0003
SIZE	1	0.02557	0.01705	1.50	0.1438
HP	1	-0.14873	0.03608	-4.12	0.0003
WEIGHT	1	-2.97682	2.72964	-1.09	0.2839
ACCEL	1	-0.69925	0.23397	-2.99	0.0054
ENG_TYPE	1	5.80842	1.56651	3.71	0.0008

Figure 8 Imputation 3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1472.16820	245.36137	66.77	<.0001
Error	31	113.92259	3.67492		
Corrected Total	37	1586.09079			

Root MSE	1.91701	**R-Square**	0.9282
Dependent Mean	24.76053	**Adj R-Sq**	0.9143
Coeff Var	7.74220		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	68.55240	4.11623	16.65	<.0001
CYLINDERS	1	-2.85923	0.76768	-3.72	0.0008
SIZE	1	0.04358	0.01883	2.31	0.0274
HP	1	-0.15208	0.04107	-3.70	0.0008
WEIGHT	1	-4.65964	2.80249	-1.66	0.1065
ACCEL	1	-0.57066	0.27764	-2.06	0.0483
ENG_TYPE	1	5.19245	1.39093	3.73	0.0008

Figure 9 Imputation 4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1460.37396	243.39566	60.02	<.0001
Error	31	125.71683	4.05538		
Corrected Total	37	1586.09079			

Root MSE	2.01380	**R-Square**	0.9207
Dependent Mean	24.76053	**Adj R-Sq**	0.9054
Coeff Var	8.13310		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	67.01166	4.22996	15.84	<.0001
CYLINDERS	1	-2.69887	0.81016	-3.33	0.0022
SIZE	1	0.04106	0.01801	2.28	0.0296
HP	1	-0.13697	0.03464	-3.95	0.0004
WEIGHT	1	-6.12984	2.49776	-2.45	0.0199
ACCEL	1	-0.35255	0.26993	-1.31	0.2011
ENG_TYPE	1	6.29941	1.72013	3.66	0.0009

Figure 10 Imputation 5

Step 3 – Combine Analysis Results

This step averages the results from the multiple analyses.

The MIANALYZE Procedure

Model Information	
Data Set	WORK.OUTREG
Number of Imputations	5

Variance Information (5 Imputations)							
	Variance						
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
cylinders	0.015726	0.568977	0.587847	3881.6	0.033166	0.032600	0.993522
size	0.000112	0.000335	0.000469	48.531	0.402702	0.314759	0.940776
hp	0.000533	0.001484	0.002124	44.079	0.431107	0.330924	0.937924
weight	5.176341	7.777031	13.988640	20.286	0.798712	0.491796	0.910449
accel	0.079946	0.072036	0.167971	12.262	1.331762	0.627338	0.888520
eng_type	0.197465	2.258081	2.495039	443.48	0.104938	0.099026	0.980579
Intercept	4.645655	16.292645	21.867431	61.546	0.342166	0.278022	0.947325

Figure 11 Combined Analysis

Parameter Estimates (5 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
cylinders	-2.892369	0.766712	-4.3956	-1.38917	3881.6	-3.037374	-2.698869	0	-3.77	0.0002
size	0.030931	0.021663	-0.0126	0.07447	48.531	0.020523	0.043585	0	1.43	0.1597
hp	-0.158924	0.046085	-0.2518	-0.06605	44.079	-0.197652	-0.136972	0	-3.45	0.0013
weight	-3.214728	3.740139	-11.0095	4.58001	20.286	-6.129836	-0.268453	0	-0.86	0.4001
accel	-0.721966	0.409842	-1.6128	0.16889	12.262	-1.071906	-0.352546	0	-1.76	0.1030
eng_type	5.855301	1.579569	2.7509	8.95967	443.48	5.192451	6.299408	0	3.71	0.0002
Intercept	69.542738	4.676262	60.1937	78.89183	61.546	67.011662	72.391093	0	14.87	<.0001

Figure 12 Combined Analysis

CONCLUSION

	Original	Original	Paramter Estimates using 5 datasets	
	Estimate	Std Error	Combined Estimate	Combined Std Error
Intercept	70.14772	8.03838	69.542738	4.676262
cylinders	-3.33403	1.56072	-2.892369	0.766712
size	0.0228	0.03207	0.030931	0.021663
hp	-0.19546	0.08065	-0.158924	0.046085
weight	-0.30623	5.13263	-3.214728	3.740139
accel	-0.78199	0.58264	-0.721966	0.409842
eng_type	6.5988	3.59008	5.855301	1.579569

Figure 13 Combined Results

In all cases, the coefficients of the combined analysis estimates gave us more precise (with less standard error) and with better p-values (< .05). The multiple imputation method MCMC greatly improved the estimates and also combining the normally distributed imputations gave us better results.

REFERENCES

Allison P. (2012). Why You Probably Need More Imputations Than You Think. http://www.statisticalhorizons.com/more-imputations. Retrieved 11/10/2013.

http://home.cerge-ei.cz/mittag/papers/Imputations.pdf

https://support.sas.com/resources/papers/proceedings11/351-2011.pdf

http://www.stat.columbia.edu/~gelman/arm/missing.pdf

https://communities.sas.com/t5/SAS-Statistical-Procedures/Maximum-Likelihood-Estimation-or-OLS-PROC-REG-vs-PROC-MIXED/td-p/197429

http://www.biostat.umn.edu/~john-c/5421/notes.019

https://statisticalhorizons.com/ml-better-than-mi

http://www.theanalysisfactor.com/missing-data-two-recommended-solutions/

http://www2.sas.com/proceedings/sugi30/111-30.pdf

http://home.cerge-ei.cz/mittag/papers/Imputations.pdf