Case Study number 1

MSDS 7333

Scott Anderwald

Introduction:

Nearly all datasets have some amount of missing observations. Several methods exist to handle the missing observations. This case study provides a means to evaluate two methods. These include listwise deletion and imputation methods. Listwise deletion is the standard method of many software packages. With listwise deletion observations are deleted which if the dataset is large enough could potentially lead to an inaccurate analysis of the data. Imputation methods substitute the missing values calculated from existing values. The purpose of the case study is to determine if using methods beyond listwise deletion improves the analysis.

Background:

The data set for the case study appears to be originally from the Carnegie Mellon University Stats Lab Library. Which was from the 1983 American Statistical Association Exposition. Within the dataset, there are nine attributes with 38 observations. The nine attributes include the following: Auto (including make and model), MPG, Cylinders, Size, HP, Weight, Accel, and Eng_Type. See table 1. For this study, the response variable will be MPG (miles per gallon).

Table1.

Obs.	Auto	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE
1	Buick Estate Wagon	16.9	8	350	155	4.36	14.9	1
2	Ford Country Sq. Wagon	15.5	8	351	*	4.054	14.3	1
3	Chevy Malibu Wagon	19.2	8	267	125	3.605	15	1
4	Chrys Lebaron Wagon	18.5	8	360	130	3.94	13	1
5	Chevette	30	4	98	68	2.155	16.5	0

The imputation method works by substituting missing values from calculations of existing variables. Values from the multiple imputations are drawn from a distribution which will contain some values. After performing the multiple imputation standard statistical analysis can be performed where the results are combined for the analysis. For this case study, a linear regression will be performed before and after the PROC MCMC step. Three steps are required for the multiple imputations. First datasets are created usually 5- 10 sets from the imputation algorithm. Second, analyze the imputed data sets with no missing values. And third, combine the analysis results by utilizing PROC MIANALYZE.

Methods:

For comparison, the data set was first analyzed by utilizing a linear regression method. With SAS being a typical software package listwise deletion was automatically performed on the dataset. Post analysis revealed that 20 out of the 38 observations had missing values. This would leave only 18 observations used for the model. see table 2.

Table. 2

Number of Observation	38
Number of Observation Used	18
Number of Observations with Missing values	20

A determination is needed to understand the mechanism for missing observations. These mechanisms include: MCAR (missing completely at random) which is the probability of Yi is not related to the value of Yj. MAR (missing at random) the probability of Y is unrelated to the value of Y controlling for another variable. And, NMAR (missing not at random) missing values that depend on observed values

Since it was noted only 18 observations were used for the linear regression analysis a determination of the missing value pattern was needed. SAS PROC MI can be used to determine the pattern of missing values see table 3 below.

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	47.37	26.605556	5.333333	177.055556	101.888889	2.795333	14.355556	0.333333
2	X	X	X	X	X	X	.	2	5.26	31.350000	4.000000	95.000000	70.000000	2.125000	16.850000	.
3	X	X	X	X	X	.	X	1	2.63	18.200000	8.000000	318.000000	135.000000	3.830000	.	1.000000
4	X	X	X	X	X	.	.	1	2.63	17.600000	8.000000	302.000000	129.000000	3.725000	.	.
5	X	X	X	X	.	X	X	3	7.89	28.133333	4.666667	128.000000	72.666667	.	16.166667	0
6	X	X	X	X	.	.	X	1	2.63	21.500000	4.000000	121.000000	110.000000	.	.	0
7	X	X	X	.	X	X	X	5	13.16	22.320000	5.400000	182.800000	.	3.009800	15.240000	0.400000
8	X	X	.	X	X	X	X	2	5.26	19.100000	6.000000	.	115.000000	3.112500	15.150000	0
9	X	X	.	X	.	X	X	1	2.63	30.500000	4.000000	.	78.000000	.	14.100000	0
10	X	.	X	X	X	X	X	2	5.26	21.100000	.	176.000000	110.000000	3.087500	15.750000	0
11	X	.	X	X	X	.	X	1	2.63	18.100000	.	258.000000	120.000000	3.410000	.	0
12	X	.	X	X	.	X	X	1	2.63	17.000000	.	305.000000	130.000000	.	15.400000	1.000000

From the pattern, it appears from the dataset can be classified as arbitrary in nature. Group one has the highest frequency for the dataset which validates the number of observations used in the linear

regression analysis. Determination of what method to be used going forward can be seen from the following table.

Pattern of Missingness	Type of Imputed Variable	Type of Covariates	Available Methods
Monotone	Continuous	Arbitrary	• Monotone regression
			• Monotone predicted mean matching
			• Monotone propensity score
Monotone	Classification (ordinal)	Arbitrary	• Monotone logistic regression
Monotone	Classification (nominal)	Arbitrary	• Monotone discriminant function
Arbitrary	Continuous	Continuous	• MCMC full-data imputation
			• MCMC monotone-data imputation
Arbitrary	Continuous	Arbitrary	• FCS regression
			• FCS predicted mean matching
Arbitrary	Classification (ordinal)	Arbitrary	• FCS logistic regression
Arbitrary	Classification (nominal)	Arbitrary	• FCS discriminant function

Source of table:

support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect019.htm

In this case study, the pattern of missingness is arbitrary and with many of the missing values being continuous. Only ENG_TYPE could be ordinal in nature. The method used for the case study was MCMC full-data imputation. From the Proc MCMC analysis, there were 25 imputations versus to typical 5-10. The system itself selected the 25 one would need to wonder if due to the fact only 38 observations were in the dataset. See table below for the output.

Model Information	
Data Set	WORK.CAR
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	501213

The MCMC algorithms works by creating "multiple imputations by using simulations from a Bayesian prediction distribution for normal data. The MCMC procedure imputes enough values to make the missing data appear monotone" (source https://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf)

While it typical for the system to use 5-10 imputations for the analysis SAS chose 25. It could be since a few observations dictated the change in imputations from the typical 5-10. 200 burn-in iterations were used before the first imputation and 100 iterations between imputations. "The burn-in iterations are used to make the iterations converge to the stationary distribution before the imputation.

The expectation maximization (EM) algorithm is a technique that finds maximum likelihood estimates for parametric models for incomplete data" ("Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0), Yuan, Yang, SAS Institute)

After running the PROC MCMC method the output is analyzed once again by the linear regression method. To recall, the pre-MCMC run resulted in only 18 out of the 38 observations being utilized in

the linear regression model. The post-PROC MCMC analysis the observation used for the linear regression analysis increased from 18 to 38. See table below

Linear Regression Post PROC MCMC

Number of observations Read	38
Number of observations Used	38

The final step in the case study is to compare the results from the initial linear regression to one that utilizes the 38 observations. The original analysis and combined are displayed in the table below.

The notable difference can be seen the standard error. For example, the Intercept parameter Std. Error improved from an initial value of 8.038 to 4.60. Eng_Type which appears to be ordinal in nature improved from 3.59 to 1.78.

Parameter	Original Estimate	Original Std. Error	Combined Estimate	Combined Std. Error
Intercept	70.14772	8.03838	70.710895	4.600770
Cylinders	-3.33403	1.56072	-3.026647	0.841981
Size	0.02280	0.03207	0.028523	0.022559
HP	-0.19546	0.08065	-0.175344	0.047250
Weight	-0.30623	5.13263	-2.125369	3.578762
Accel	-0.78199	0.58264	0.817424	0.325886
Eng_Type	6.59880	3.59008	5.831644	1.782942

By using PROC MCMC with the default settings it improved the linear analysis output.

Conclusion:

Like many datasets in the world, the case study had several missing observations. Dealing with observations can be undertaken by several methods which include listwise deletion and imputations. For this case study, multiple imputations were used to improve the linear regression analysis. By comparing the analysis there is a notable improvement in the parameters. One should note that the method chosen should be made by examining the individual datasets and determine which will work the best.