Multiple Imputations

James Tsai, Wid Sogata

MSDS 7333 - Quantifying the World, 01/26/17

ABSTRACT

Missing data is a very common issue in a high number of research fields. In many cases, this situation undermines the accuracy and validity of the results. Multiple Imputation (MI) is a statistical technique that can be used to improve this condition by filling in the missing values. Its advantage compared to other techniques such as single imputation or complete case, is in its flexibility. This technique can be used in cases where the data is missing at random, completely at random or missing not at random.

This paper reviews this method of analyzing missing data and the application of MI techniques by averaging the outcomes developed by Donald B. Rubin (1987), Harvard University. At the end a comparison of regression results between list-wise deletion and multiple imputation is discussed.

Keywords: missing data, multiple imputations, methods, validity, accuracy, iterations, list-wise deletion, complete case, SAS, MI.

INTRODUCTION

This paper examines MI method of analyzing missing data and application of it by averaging the outcomes of several iterations of imputation. Rubin developed this technique by replacing each missing data with a set of probable replacements that represent the uncertainty about the right value to input. Using a data set, we demonstrate the comparison of using list-wise deletion (complete case) with MI. By maximizing the use of available data, we intend to show the benefit and improvement of utilizing MI outweighs the simplicity of the former, especially in the case of small number of data with complete cases.

BACKGROUND

The software we are going to use is SAS version 9.4, which includes MI (Multiple Imputation) procedure for creating multiple imputations for incomplete data as well as producing analysis resulting from calculation of multiple imputed data sets.

This data set consists of technical specifications of multiple brands of cars in the world. Listed in Table 1 are the variable names, description, and attribute type from the data set.

Variable Name	**Description**	**Attribute Type**
auto	Brand of the car	Unique string

mpg	Miles per gallon consumption	Continuous
cylinders	Number of cylinders	Multi-valued discrete
Size	Displacement	Continuous
hp	Horsepower	Continuous
weight	Weight of the car	Continuous
accel	0 to 60 mph acceleration	Continuous
eng_type	Engine type	Multi-valued discrete

Table 1. Data Attributes

We are interested in creating a model to predict the value of mpg through linear regression method given the variables cylinders, size, horsepower and weight. We will leave out acceleration and engine types from this model.

METHODS

Data Analysis

The data consists of comparison of cars technical specifications that includes multiple brands in the world. There are 38 observations with 8 attributes of each. Within the data there are 16 observations with missing values. Based on Table 2, the pattern of missing data is non-monotone (arbitrary) since we cannot reorder the variable to make it monotone.

Group	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE	Freq	Percent	Group Means						
										MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE
1	X	X	X	X	X	X	X	18	47.37	26.61	5.33	177.06	101.89	2.80	14.36	0.33
2	X	X	X	X	X	X	.	2	5.26	31.35	4.00	95.00	70.00	2.13	16.85	.
3	X	X	X	X	X	.	X	1	2.63	18.20	8.00	318.00	135.00	3.83	.	1
4	X	X	X	X	X	.	.	1	2.63	17.60	8.00	302.00	129.00	3.73	.	.
5	X	X	X	X	.	X	X	3	7.89	28.13	4.67	128.00	72.67	.	16.17	0
6	X	X	X	X	.	.	X	1	2.63	21.50	4.00	121.00	110.00	.	.	0
7	X	X	X	.	X	X	X	5	13.16	22.32	5.40	182.80	.	3.01	15.24	0.4
8	X	X	.	X	X	X	X	2	5.26	19.10	6.00	.	115.00	3.11	15.15	0
9	X	X	.	X	.	X	X	1	2.63	30.50	4.00	.	78.00	.	14.10	0
10	X	.	X	X	X	X	X	2	5.26	21.10	.	176.00	110.00	3.09	15.75	0
11	X	.	X	X	X	.	X	1	2.63	18.10	.	258.00	120.00	3.41	.	0
12	X	.	X	X	.	X	X	1	2.63	17.00	.	305.00	130.00	.	15.40	1

Table 2. Missing Data Patterns

Model Fit / Diagnostic Plot

We continue our data analysis with model fit verification for linear regression. Figure 1 shows the result of the observations.

Residual Plot: Residual Plot: The residual plot resembles a random scatter of data points around the 0. This indicates a linear regression model is appropriate for the data.
Quantile-Quantile Plot of Residuals: The QQ Plot of residuals displayed provides no evidence that the residuals are not normally distributed.
Cook's Distance: This influence indicator shows there is one extreme outlier (observation 10), which should not be a concern.
Studentized Residual Plot: This plot offers an alternative criterion to identify outliers. The result is similar to the random pattern to the Residual plot. This plot is also consistent with Cook's Distance plot, which shows one extreme outlier.
Histogram of Residuals: The histogram of residuals displayed in the table does not provide strong evidence that the residuals are not normally distributed.

Based on the diagnostic plots, it is reasonable to proceed with the Linear Regression model for mpg prediction.

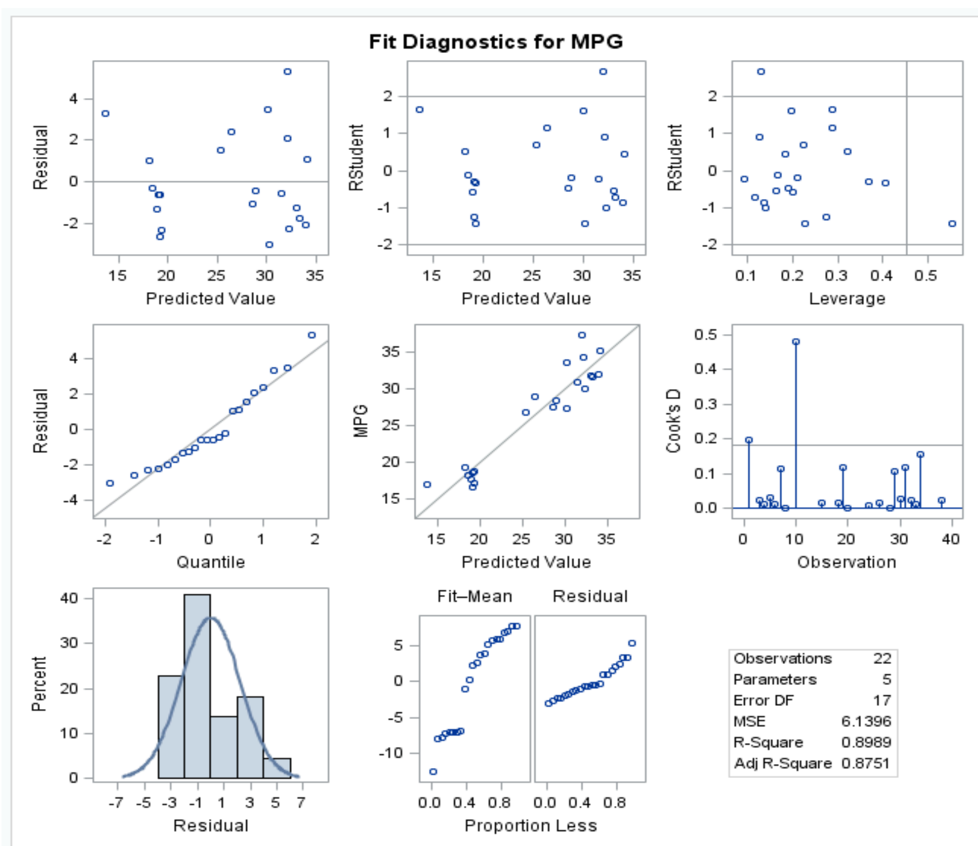

Figure 1. Diagnostic Plot for MPG

Results of the Initial Regression Analysis

As we can see from the table 3, the corrected total degree of freedom is only 21 since we are not utilizing the whole data set due to default list-wise deletion in SAS. Effectively, we have reduced the statistical power in our regression analysis when running the linear regression procedure (PROC REG).

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	927.64081	231.9102	37.77	<.0001
Error	17	104.37374	6.13963		
Corrected Total	**21**	1032.0146			

Table 3. Analysis of Variance (Complete Case Analysis)

Table 4 shows our model parameter results with standard error estimation using the complete case method.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	**59.29187**	**4.60156**	12.89	<.0001
CYLINDERS	**1**	**-1.52024**	**1.06901**	-1.42	0.1731
SIZE	**1**	**0.06595**	**0.02756**	2.39	0.0285
HP	**1**	**-0.06502**	**0.05948**	-1.09	0.2895
WEIGHT	**1**	**-10.66719**	**3.0213**	-3.53	0.0026

Table 4. Parameter Estimates (Complete Case Analysis)

Multiple Imputation

Before continuing with the multiple imputations, we must first understand the conditions in which it makes sense to apply multiple imputation technique. The data must satisfy one of two conditions. It must either be Missing Completely at Random (MCAR) or Missing at Random (MAR). If the data is Missing Not at Random (MNAR), it will not be suitable for running any MI technique.

Specifically, the SAS imputation technique assumes that the probability of a missing observation may be dependent on an observed value $\mathbf{Y}_{\mathbf{obs}}$, but not on $\mathbf{Y}_{\mathbf{mis}}$ (Rubin 1987, p. 53). When the missing value(s) has a dependency on the observed values, this is known as MAR. In the MCAR case, there is no dependency on observed values, and this can be thought of as a special case of MAR.

We also assume that the missing data are from a continuous multivariate distribution and contain missing values that can happen to any variables in the data set. It also assumes that the existing data are from normal distribution when regression method is used.

There are three methods that are available in the MI procedure. The method chosen depends on the type of missing data pattern. For monotone missing data patterns, either a parametric regression method that assumes multivariate normality or a nonparametric method that uses propensity scores is appropriate. For an arbitrary missing data pattern, a Markov Chain Monte Carlo (MCMC) method (Schafer 1997) that assumes multivariate normality can be used.

Table 5 shows model setup for MI procedure in SAS.

Model Information	
Data Set	WORK.CARS_DATA
Method	**MCMC**
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	0
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	689640001

Table 5. Model Information of MI

The folowing table illustrates the removal of observations (2, 8, 9, 11, 12) from the first 12 observations due to the list-wise deletion and the corresponding observations in the multiple imputations data set where all the observations are kept.

List-wise Deletion								
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	150	3.94	13	1
5	Chevette	30	4	98	68	2.155	16.5	0
6	Toyota Corona	27.5	4	134	95	2.56	14.2	0
7	Datsun 510	27.2	4	119	97	2.3	14.7	0
8	~~Dodge Omni~~	~~30.9~~	~~4~~	~~105~~	~~75~~	~~2.23~~	~~14.5~~	
9	~~Audi 5000~~	~~20.3~~	~~5~~	~~131~~	~~.~~	~~2.83~~	~~15.9~~	0
10	Volvo 240 GL	17	6	163	125	3.14	13.6	0
11	~~Saab 99 GLE~~	~~21.6~~	~~.~~	~~121~~	~~115~~	~~2.795~~	~~15.7~~	~~0~~
12	~~Peugeot 694 SL~~	~~16.2~~	~~6~~	~~.~~	~~133~~	~~3.41~~	~~15.8~~	~~0~~

Multiple Imputation

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	**135.55474**	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	150	3.94	13	1
5	Chevette	30	4	98	68	2.155	16.5	0
6	Toyota Corona	27.5	4	134	95	2.56	14.2	0
7	Datsun 510	27.2	4	119	97	2.3	14.7	0
8	Dodge Omni	30.9	4	105	75	2.23	14.5	**-0.0630776**
9	Audi 5000	20.3	5	131	**98.552396**	2.83	15.9	0
10	Volvo 240 GL	17	6	163	125	3.14	13.6	0
11	Saab 99 GLE	21.6	**4.019347**	121	115	2.795	15.7	0

12	Peugeot 694 SL	16.2	6	**206.26745**	133	3.41	15.8	0

Table 6. Comparison of Complete Case and Multiple Imputed Data Set Treatments

Results of the Regression Analysis after Imputation Process

The default Method is MCMC since the missing data pattern is arbitrary. The number of imputation for data set created is defaulted to 25. We noticed that SAS recently changed default value from 5 to 25. This improvement is appropriate in many situations especially when the missing values are in high percentage. It is recommended to have 20 imputations for 10% to 30% missing data and 40 imputations for 50% missing data (Graham, 2008). In terms of complete cases, we were missing approximately 41% of the data. Therefore, we decided to keep the default at 25, which seems appropriate.

Table 7 shows the first imputation result. Notice that the corrected total degree of freedom is now 37, which means that the model uses all 38 observations and increased statistical power.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	1383.44415	345.86104	56.32	<.0001
Error	33	202.64664	6.14081		
Corrected Total	**37**	1586.09079			

Table 7. Analysis of Variance of First Imputation

Table 8 shows the first 3 parameters results of regression procedure from the each of the 25 imputation iterations.

Predicting MPG												
Obs	**_Imputation_**	**_MODEL_**	**_TYPE_**	**_NAME_**	**_DEPVAR_**	**_RMSE_**	**Intercept**	**CYLINDERS**	**SIZE**	**HP**	**WEIGHT**	**MPG**
1	**1**	**MODEL1**	**PARMS**		**MPG**	**2.47807**	**60.24**	**-1.29832**	**0.071834**	**-0.0459**	**-12.795**	**-1**
2	1	MODEL1	COV	Intercept	MPG	2.47807	12.96	-1.44508	0.065953	0.0042	-6.0369	.
3	1	MODEL1	COV	CYLINDERS	MPG	2.47807	-1.445	0.68044	-0.01113	-0.0036	0.0534	.
4	1	MODEL1	COV	SIZE	MPG	2.47807	0.066	-0.01113	0.000441	9.4E-05	-0.0331	.
5	1	MODEL1	COV	HP	MPG	2.47807	0.004	-0.00355	0.000094	0.00154	-0.0556	.
6	1	MODEL1	COV	WEIGHT	MPG	2.47807	-6.037	0.05343	-0.03313	-0.0556	6.066	.
7	**2**	**MODEL1**	**PARMS**		**MPG**	**2.48353**	**59.36**	**-1.67484**	**0.070604**	**-0.0239**	**-12.498**	**-1**

8	2	MODEL1	COV	Intercept	MPG	2.48353	11.92	-1.27692	0.059456	0.01403	-5.9239	.
9	2	MODEL1	COV	CYLINDERS	MPG	2.48353	-1.277	0.60689	-0.00944	-0.0045	0.0434	.
10	2	MODEL1	COV	SIZE	MPG	2.48353	0.06	-0.00943	0.000394	0.00016	-0.0334	.
11	2	MODEL1	COV	HP	MPG	2.48353	0.014	-0.00448	0.000163	0.00144	-0.059	.
12	2	MODEL1	COV	WEIGHT	MPG	2.48353	-5.924	0.04336	-0.03345	-0.059	6.2395	.
13	**3**	**MODEL1**	**PARMS**		**MPG**	**2.60641**	**60.72**	**-0.78556**	**0.070532**	**-0.0437**	**-13.959**	**-1**
14	3	MODEL1	COV	Intercept	MPG	2.60641	16.46	-1.40017	0.080501	-0.0222	-7.3401	.
15	3	MODEL1	COV	CYLINDERS	MPG	2.60641	-1.4	0.59473	-0.00977	-0.0012	0.038	.
16	3	MODEL1	COV	SIZE	MPG	2.60641	0.081	-0.00976	0.000496	-0.0001	-0.0358	.
17	3	MODEL1	COV	HP	MPG	2.60641	-0.022	-0.00122	-0.00015	0.00127	-0.0255	.
18	3	MODEL1	COV	WEIGHT	MPG	2.60641	-7.34	0.03799	-0.03576	-0.0255	5.6388	.

Table 8. The First 3 Imputations Model Parameters

RESULTS

As a conclusion of this exercise, we observe using SAS MIANALYZE procedure combined results across imputations. Regression coefficients are averaged across imputations. Uncertainties from 2 sources were incorporated to produce the standard errors. These were "within" imputation and "between" imputation. "Within" imputation is the estimate variability expected with completed data by imputation. "Between" imputation is the estimate variability due to missing information or uncertainty surrounding missing values.

Furthermore, we were able to achieve more confidence in our conclusions with a bigger statistical power. In this particular case we see a smaller standard error, resulting in greater accuracy in our model.

Parameter Estimates (25 Imputations)										
Parameter	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**Minimum**	**Maximum**	**Theta0**	**t for H0: Parameter= Theta0**	**Pr > |t|**
Intercept	59.872578	**3.68093**	52.6564	67.08877	5140.6	58.019698	61.58232	0	16.27	<.0001
cylinders	-1.31651	**0.850487**	-2.9846	0.35161	1691.7	-2.052196	-0.785562	0	-1.55	0.1218
size	0.068822	**0.020717**	0.0282	0.10943	7398.5	0.054244	0.07497	0	3.32	0.0009
hp	-0.050319	**0.041224**	-0.1312	0.03053	1705.5	-0.0843	-0.023756	0	-1.22	0.2224
weight	-12.28264	**2.574687**	-17.3312	-7.23411	2731.9	-13.959117	-9.990991	0	-4.77	<.0001

Table 9. Parameter Estimate of Combine Imputations

Table 10 shows comparison of the initial regression to the second regression using combined imputation data. We noticed smaller standard error for each parameter in the model.

Parameter Estimates					
		Initial Parameter		After Multiple Imputation	
Variable	DF	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	**1**	59.29187	**4.60156**	59.8726	**3.68093**
CYLINDERS	**1**	-1.52024	**1.06901**	-1.3165	**0.85049**
SIZE	**1**	0.06595	**0.02756**	0.06882	**0.02072**
HP	**1**	-0.06502	**0.05948**	-0.0503	**0.04122**
WEIGHT	**1**	-10.66719	**3.0213**	-12.283	**2.57469**

Table 10. Comparison of Initial Parameters and Combined Imputations Parameters Result

At the end, our final model for MPG is:

$$MPG = 59.29 - (1.52)\ CYLINDERS + (0.66)\ SIZE - (0.065)\ HP - (10.67)\ Weight$$

CONCLUSION

By using MI technique, it can provide an analysis that account for the uncertainty due to missing values. By generating a random representative sample of the missing values, it maximizes the use of available data. In many cases it produces better and more accurate results by reducing bias and producing appropriate estimates of uncertainty, which would not be obtained otherwise.

REFERENCES

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Cambridge, MA: John Wiley & Sons, Inc.

Yuan, Y.C. *Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0).* Rockfield, MD: SAS Institute, Inc.

Allison, P. *Why You Probably Need More Imputations Than You Think.* (2012). Retrieved January 20, 2017, from http://statisticalhorizons.com/more-imputations.

Bodner, T. E. *What Improved with Increased Missing Data Imputations?* (2008). Structural Equation Modeling: A Multidisciplinary Journal.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Graham, J.W. (2008) *Missing Data Analysis: Making It Work in the Real World*. University Park, PA.