University of California
Los Angeles

Missing Data Imputation

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Ling Jing

2012

Abstract of the Thesis

Missing Data Imputation

by

Ling Jing

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Ying Nian Wu, Chair

Existence of missing values creates a big problem in real world data. Unless those values are missing completely at random, we cannot disregard them. This paper demonstrates some implementation methods to deal with missing values. It will show the theory and implementations of EM Algorithm, Regression Imputation, Stochastic Regression Imputation and Multiple imputations. This paper begins by introducing the theories of those methods and then applying them to two examples and finally diagnostics. I will use some examples to demonstrate these algorithms and hope this helps researchers with various backgrounds solve missing values problem in their data.

The thesis of Ling Jing is approved.

Nicolas Christou

Jan de Leeuw

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2012

To my mom

for her love and support.

And to my teachers, classmates and Statistics Department staff

who have given me precious memories and tremendous encouragement.

Table of Contents

List of Figures . . . . . **vii**

List of Tables . . . . . **viii**

1 Introduction **1**

2 Imputation Methods **3**

2.1 EM Algorithm 3

2.1.1 Theory 3

2.1.2 Imputations 5

2.2 Regression Imputations 5

2.2.1 Routine multivariate imputation 6

2.2.2 Iterative Regression Imputation 6

2.2.3 Stochastic Regression Imputation 7

2.3 Multiple Imputations 7

3 Examples **9**

3.1 Airport Delay Data 9

3.2 American time usage survey Data 9

4 Procedures **11**

4.1 Algorithm Preparation 11

4.2 Data Preparation 12

5 Results and Diagnostics 13

5.1 Visualization and Diagnostics for Multiple Imputations 13

5.2 Diagnostics 16

5.2.1 Correlations 17

5.2.2 Difference in Means 19

5.2.3 Confidence Intervals Vs. Real Mean 20

6 Discussions and Conclusions 29

A Appendix A. Data Explanation 31

Bibliography 32

List of Figures

5.1 Missing values distribution graph, Red indicates missings in the data 14

5.2 Example of Diagnostics graph on Airport Delay Data 15

5.3 Example of Diagnostics graph on American Time Usage Data 16

List of Tables

5.1 Diagnostics: Correlation Comparison Table on Airport Delay data . 17

5.2 Diagnostics: Correlation Comparison Table on American Time Usage 18

5.3 Diagnostics: Difference between real and predicted means Table on Airport data 19

5.4 Diagnostics: Difference between real and predicted means Table on American Time Usage data 20

5.5 Diagnostics: 95% Confidence Interval for imputed values for EM on Airport Data 21

5.6 Diagnostics: 95% Confidence Interval for imputed values for Regression Imputation on Airport Data 22

5.7 Diagnostics: 95% Confidence Interval for imputed values for Sto. Regression Imputation on Airport Data 23

5.8 Diagnostics: 95% Confidence Interval for imputed values for MI on Airport Data 24

5.9 Diagnostics: 95% Confidence Interval for imputed values for EM on American Time Usage Data 25

5.10 Diagnostics: 95% Confidence Interval for imputed values for Regression Imputations on American Time Usage Data 26

5.11 Diagnostics: 95% Confidence Interval for imputed values for Stoc. Regression Imputation on American Time Usage Data 27

5.12 Diagnostics: 95% Confidence Interval for imputed values for MI on American Time Usage Data 28

Acknowledgments

I would like to acknowledge the advice from UCLA lecturer Vivian Lew for giving me the idea of this thesis; also, I would like to thank my friend Yao Shi's help on debugging my LaTeX code. And also my classmate and friend Matthew Baltar for editing my paper.

CHAPTER 1

Introduction

Existence of missing values is common in real world data. If the occurrence of the missing data is completely at random, we can simply remove it. In many cases missing values happen when there is a human mistake when collecting the data or some information is damaged and it will make the data biased if we remove those missing observations. With statistical tools, we can impute the missing values so the maximum amount of information is restored while keeping the data unbiased. This paper demonstrates some of the popular statistical methods for imputing missing values. With the information restored, it can help us in performing better analysis and making better business decisions. Some of the common methods in data imputation, which will be demonstrated in this paper, include EM Algorithm, Regression imputation, Stochastic Regression Imputation and Multiple imputation.

In any research, if we encounter missing value problems, the first thing we should do is to examine the natural pattern of the missing values. According to Little and Rubin (1987), missing patterns generally fall into one of three types, "Missing Completely at Random", "Missing at Random" and "Not Missing at Random". To distinguish which category the data fall into, we can calculate the percentage of missing values for each demographic category. For example, if a set of "household income" values are missing, we want to look at the percentage missing within each ethnicity group and gender and age. If all percentages of missing are the same or similar for all individual groups, this is called "Missing Completely at Random"(MCAR), and

the missing observations can be safely removed.

If the percentage of missing is the same or similar for some groups and not for all, then it is "Missing at Random"(MAR). e.g. percentage of missing are the same across gender and ethnicity but not age. In this case, we can use many different methods to impute the missing values, and kind of data is the objective of this thesis.

The third case is the most difficult situation to deal with. When data is censored it is a "Not Missing at Random"(NMAR) or "nonignorable", and this is a damaging situation. To help people understand this situation, I am going to use Little and Rubin's example shown in their book called Statistical Analysis with Missing Data. If a variable measures time to the occurence of an event such as death of an experimental animal; some units in the experiment are terminated before the event occurs. If the time to censoring or termination is known, we at least know the time of occurence passes that time. The analysis needs to take account of this information to avoid biased conclusions [3].

CHAPTER 2

Imputation Methods

2.1 EM Algorithm

2.1.1 Theory

Expectation Maximization algorithm was introduced to the public in 1977. EM algorithm is commonly used in data clustering and Machine learning. According to Rubin, The EM algorithm is a very general iterative algorithm for ML estimation in incomplete-data problems. The problems that can be attacked by EM algorithm include missing value estimations and imputations, variance component estimation and factor analysis[3].

EM algorithm consists of the Expectation step (E step) and Maximization step(M step). E step calculates the conditional expectation of the parameter on missing data, which are the objective values we are trying to impute given the $(t-1)$th iteration on the parameters. M step estimates the parameters by maximizing the complete data likelihood.

To express the process with math symbols, the complete-data log-likelihood is

$$l(\theta|y),$$

then mean of it is

$$Q(\theta|\theta^{(t)}) = \int l(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis},$$

and the goal of M step is

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

[3]. Until this chain reaches a point where the estimate of θ converges.

In general cases, we assume data follow a normal distribution which is the most common distribution in nature. With this assumption, we can formulate EM in a more concrete form. We know that there are two parameters that determine the difference between normal distributions, the mean μ and variance σ^2. Given $\sigma^2 = E[x^2] - (E[x])^2$, $E(x)$ and $E(x^2)$ will give us everything we need for the distribution, which are called sufficient statistics.

The deviation of EM algorithm is as the following, we assume the data is sorted so observation 1 through r is observed and r through n are missing for easy notation and explanation.
E-step estimates the value of the parameters,

$$\begin{aligned} E(y_i|\theta^{(t)}, Y_{obs}) &= \sum_{i=1}^{r} y_i + (n-r)\mu^{(t)} \\ E(y_i^2|\theta^{(t)}, Y_{obs}) &= E(y_{obs}^2 + y_{mis}^2) \\ &= E(\sum_{i=1}^{r} Y_i^2 + (n-r)(\mu^t)^2) \\ &= \sum_{i=1}^{r} Y_i^2 + E(n-r)(\mu^t)^2 \end{aligned}$$

The above estimates are called sufficient statistics; in other words, we could use these two values to sample the distribution. In this case, the first equation gives us the estimate of $(t+1)th$ step estimate μ, and the second one is the expected square sum of all observations. Since we have the formula $\sigma^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - (\frac{\sum_1^n y_i}{n})^2$, the distribution can be easily calculated with those two values.

The M-step uses the same sufficient statistics using the estimated parameters from

the E-step. The estimates are the following,

$$\begin{aligned}
\mu^{(t+1)} &= E(\sum_{i=1}^{n} Y_i|\theta^{(t)}, Y_{obs})/n \\
&= E(\sum_{i=1}^{r} Y_i + \sum_{obs=r+1}^{n} Y_{obs})/n \\
(\sigma^{(t+1)})^2 &= E(\sum_{i=1}^{n} Y_i^2|\theta^{(t)}, Y_{obs})/n - (\mu^{(t+1)})^2
\end{aligned}$$

The interpretation for $\mu^{(t+1)}$ is using the observed values and estimated parameters from E-step regarding the data as it there are no missing values then calculate the expected value. Since we know $\sigma^2 = E(X^2) - (E(X))^2$, with the t-th step estimations, $E(X^2)$ is known and so that σ^2 can be computed.

2.1.2 Imputations

Though the theory of EM is a bit complicated and hard to understand, the actual imputation is much easier. First, we start with a guessed set of values for the missing values for iteration 1. Theoretically, we can start with any values. I suggest choosing the mean of the observed values to reduce the number of iterations before its convergence. With the guessed values from step 1, we can now start step 2. Step 1 helped us finish the E-step of the algorithm, and with the M-step inference of EM Algorithm we can define the sufficient statistics of step 2. We keep repeating the steps until the $(t+1)th$ step returns the same μ and σ values as the $(t)th$ step.

2.2 Regression Imputations

This method is designed to be used for data with partially observed variable y and fully observed X variables. In simple words, this method is pretty self explanatory from its name, we run regressions on the known part of the data. With the regres-

sion model we can predict the missing values. The difficulty of this method is that it requires some effort to set up a reasonable multivariate regression model, and a better regression model produces a better result. In practice, an off-the-shelf model is typically used. Generally the multivariate normal or t distribution is used for continuous outcomes and a multinomial distribution for discrete outcomes. [1] There are a couple of different approaches to this method.

2.2.1 Routine multivariate imputation

Using the appropriate variables with complete data, we can fit a multivariate model to variables with missing values, then we use the predicted/fitted values to impute the missing. This is the most simple and direct approach to take and in practice a good place to start.

2.2.2 Iterative Regression Imputation

Iterative Regression Imputation is a different way of generating the missing values. Assuming the data has n observations, it has variable, Y, with missing values, where $y_1, ..., y_r$ are missing. Let X denote our carefully selected dependent variable in the regression and is fully observed in the data.

Step 0: Similar with EM Algorithm step 1, we start the computation with a set of guessed values $Y_1, ..., Y_r$

Step 1: Calculate $(y_1|y_2, ..., y_r, Y_{obs}, X)$

with the estimated y_1 calculate $(y_i|y_1, ..., y_{i-1}, y_{i+1}, ..., y_r, Y_{obs}, X)$ and so on.

step 2: Keep looping through the algorithm until approximate convergence.

2.2.3 Stochastic Regression Imputation

This method adds a little trick to Regression Imputation. It uses regression imputation plus a residual term to the prediction. In general case, we see the residual follows normal distribution with zero mean and the residual variance from the regression[3].

When we are using this method, we can simply add an extra sampling step to it. By adding a term sampled from the $N(0, \sigma^2)$ with regular regression imputation, we'll get the predicted values for Stochastic Regression.

2.3 Multiple Imputations

Aside from the methods stated above, there are so many more methods of imputing missing data, and all of them produce different results. Multiple imputation combines them all and then produces one overall analysis[2].

According to Gelman, Multiple imputation creates several imputed values for each missing value from similar but different methods, and each method spits out a "complete dataset". Using those datasets, we can draw a combined inference across all datasets. Gelman also gives us an example in his book. If we are using regression and we want to make inference about the coefficient $\hat{\beta}$, we can simply take the average across all datasets $\hat{\beta_m}$.

To express it in formulas,

$$
\begin{aligned}
&Mean, \\
&\qquad \hat{\beta} &= \frac{1}{m}\sum_{m=1}^{M}\hat{\beta}_m \\
&Variance, \\
&\qquad V_\beta &= \frac{1}{m}\sum_{m=1}^{M}s_m^2 + (1+\frac{1}{m})\frac{1}{m-1}(\hat{\beta}_m - \hat{\beta})^2
\end{aligned}
$$

Where m is number of methods we are using and β_m and s_m the estimates from every

individual methods.

For the actual imputations, it is rather complicated to do it from complete scratch, but there are packages within R that can be easily used. In this paper, i am using an R package called **MI**, which was developed by Yu-Sung Su and his colleagues[4], for my data.

CHAPTER 3

Examples

I have downloaded two very different data sets from public data server data.gov. Variables in $AirportDelay$ data are highly corelated, and $AmericanTimeUsage$ data are less predictive. These data are free and public, people can download it from the site if they are interested.

3.1 Airport Delay Data

The data didn't come in a nice and clean form. This thesis is trying to test those four methods mentioned in the theory section on manually removed data, so I am going to clean the data to a nice form that I can use for the imputations. Since the data contains most major cities' data in 2009, it is very large with 49,440 observations. In order to make the dataset suitable for this analysis, I used departures from LAX and observations with more than 15 min of delay time only. The sample size is then cut to 5785.

3.2 American time usage survey Data

Same as most public data, this one requires some cleaning too. I removed obsevations with demo values that didn't make sense (e.g. the interviwer's age lower than 14). I combined variable $PEHRACT1$(main job hours) and $PUHROT1$(over time hours)

together for a new variable total hours worked, and this is the variable being used as respondent for the analysis. The sample size droped from 68,846 to 1451 which is more suitable for this experiments.

CHAPTER 4

Procedures

This thesis is trying to find out what are the best approaches to take when we face different kinds of datasets with different amounts of information missing. I removed part of the data, then I put it through all four methods and analyzed the resulting data. This is a good way to test which method works the best for different percentages of missing data.

4.1 Algorithm Preparation

I am going to be using four algorithms for the data sets; EM Algorithm, Routine Multivariate Imputation, and Multiple Imputations. As I have demonstrated in the Imputation Methods chapter, I wrote code for EM and Regression imputations. Althogh EM looks complicated in theory, when we apply it to missing value imputations it is rather easy. Based on the observed distribution, we draw sample to impute the missing values one by one until (t+1)th step matches with (t)th step.

On Regression imputations, I introduced a couple of the methods in the theory part, but i am going to use the simple one, Routine Multivariate Imputation. If people who are reading this paper wants to use other regression methods, it should be easily altered to the desired approach.

Finally, I am going to use a package called **MI** in R for Multiple Imputations. It is fairly easy to use, and it takes in the data frame with NA as missing and produces

the fully imputed data that is ready for the next step.

4.2 Data Preparation

I am going to run a big number of datasets, I took out 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50% of the values in one variable for both datasets, and then made them go through all four methods, then the goal is to compare them with the original data. To eliminate noise for my experiment, I am going to remove my data randomly, so it keeps the properties of missingness consistent across all datasets. In order to find out what is the best approach for a specific dataset, I am going to make my code produce the imputed data, the index of imputed values, mean and variance of each method; thereafter, reasonable diagnostics can be performed. Since I have 20 datasets for the experiment, I need to keep the naming and coding consistent and simple. I think for people who is doing similar kind of analysis should follow the same rule too. Depending on the data size and number of missing values some algorithms (e.g. **MI**) might take a long time to finish.

Since I need a set of datasets for the experiment, a good plan is needed. First put all data into a list, name it properly and record the index of the missing values, then perform all methods in loops for fast speed. Finally, based on the info, diagnose then results.

CHAPTER 5

Results and Diagnostics

All four method ran smoothly on both of my data sets. Multiple Imputations have the longest running time, so if you are need Multiple Imputation on a large data, please reduce your data column down to a reasonable size (e.g. remove all the unnecessary variables.)

5.1 Visualization and Diagnostics for Multiple Imputations

There are not many things to visualize in EM and regression imputations, but the MI package provides a super useful tool for us to visualize the distribution of the missing values, which is really helpful for researchers to understand their data missing patterns. I would recommend everyone look at this graph before we get into the imputation precedure. I am going to show a few graphs and explain how to use it.

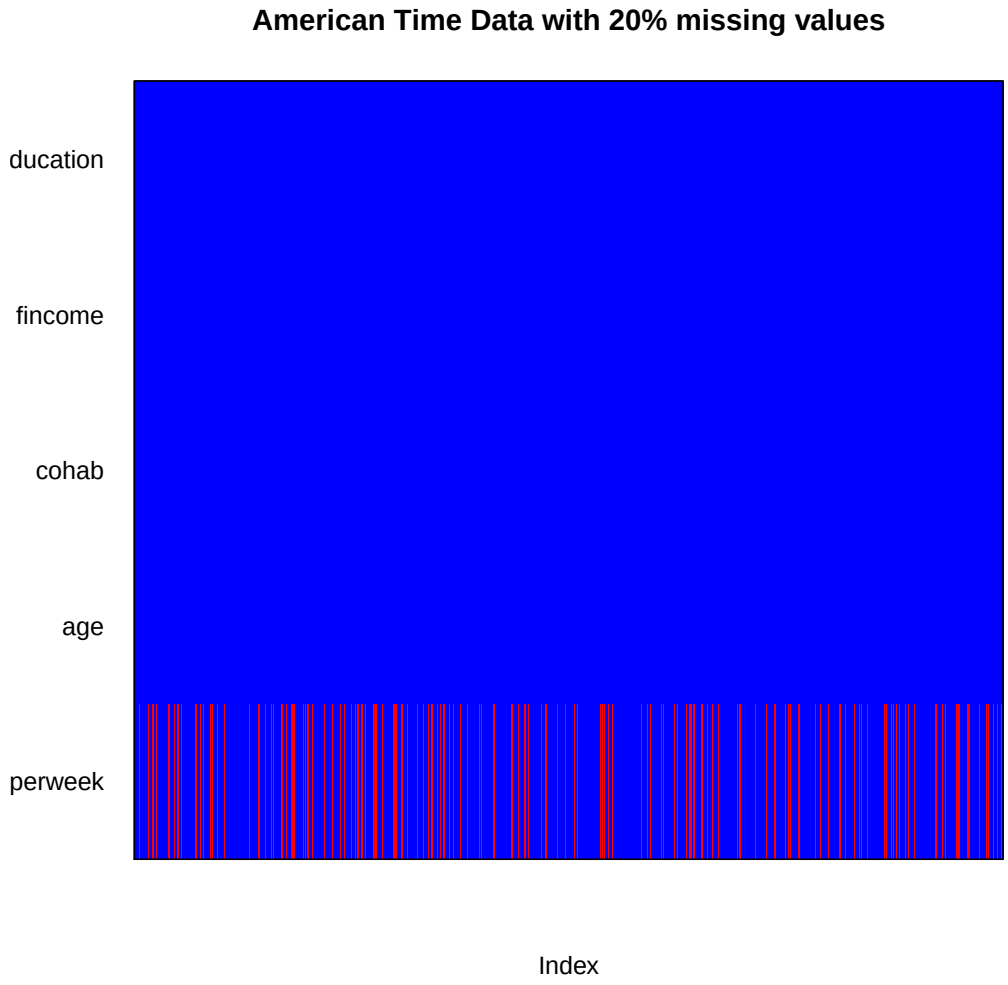

Figure 5.1: Missing values distribution graph, Red indicates missings in the data

From Figure 5.1, we can see the missing data pattern of `hrperweek` variable. The nice thing about this function is it will show you the relative missing patterns of all variables in the data, so I would recommend to only include the important variables in the data before putting it into the argument. Note, blue represent existing data and red bars shown for missing values.

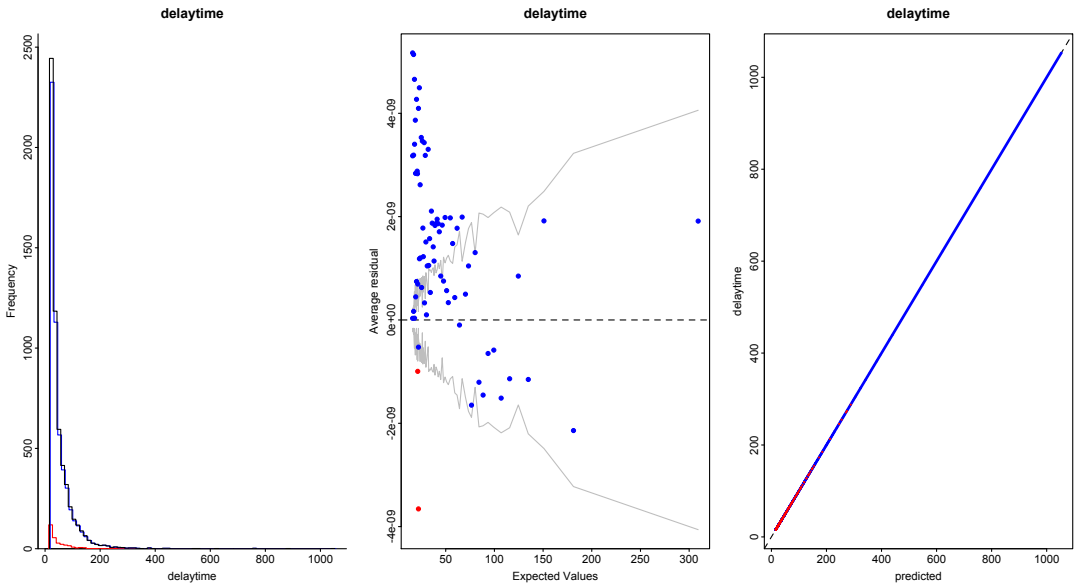

Figure 5.2: Example of Diagnostics graph on Airport Delay Data

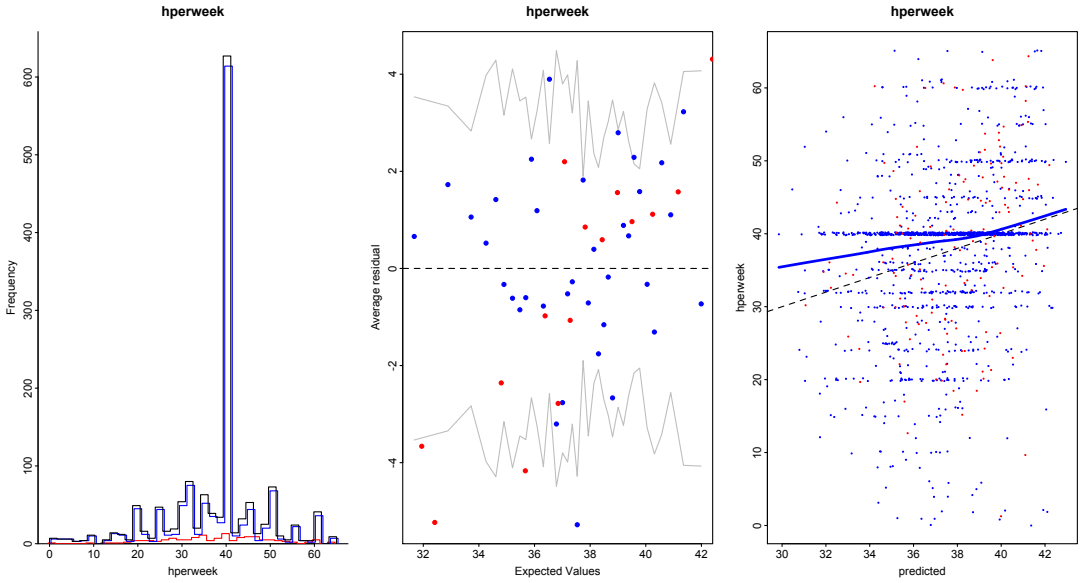

Figure 5.3: Example of Diagnostics graph on American Time Usage Data

There are 20 data sets in my experiment so I have 20 of those set of diagnostic graphs. To save space, I am only going to show 2 of them. The first one (Figure 5.2) comes from `airport delay` data, and Figure 5.3 comes from `american time usage` data.

5.2 Diagnostics

A few different methods are used in this section to diagnose the results. First, I compared the correlation of the imputed data and real data. Secondly, I used the differences in the imputed mean and real mean. Finally, I found the 95% confidence

interval of each procedure, then test if the real mean is included in the interval.

5.2.1 Correlations

A good imputation would give us values that are close to the real data and restore information close to the truth. By taking the correlations of the imputed values $\tilde{X}$, we are able to see the coherence of the imputed values and the real data. If the predicted compare well with the real data, we can say it is a good approach to take in that particular case.

Missing %	EM Algorithm	Reg. Imp.	Stoc. Reg. Imp.	MI
5%	-0.09044721	1	1	1
10%	0.0949673	1	1	1
15%	0.05876163	1	1	1
20%	-0.008840737	1	1	1
25%	0.06807379	1	1	1
30%	0.0109727	1	1	1
35%	0.007607618	1	1	1
40%	0.03902101	1	1	1
45%	0.004463893	1	1	0.9999999
50%	-0.01726519	1	1	0.9999991

Table 5.1: Diagnostics: Correlation Comparison Table on Airport Delay data

Missing %	EM Algorithm	Reg. Imp.	Stoc. Reg. Imp.	MI
5%	-0.08866691	0.2751009	0.1505338	0.2045575
10%	0.198255	0.109793	0.04266626	-2.799145e-05
15%	0.0449258	0.2015482	0.04444492	-0.004644375
20%	0.06724799	0.2679266	0.09029146	0.02241877
25%	0.03860827	0.1647267	0.03954247	0.1290694
30%	0.0274811	0.2636278	-0.02207877	0.0731526
35%	-0.02779511	0.1634827	0.02279567	0.004058042
40%	0.01282385	0.1857025	-0.05230224	0.07028629
45%	0.04876823	0.1990897	0.07142353	0.04002668
50%	0.01534002	0.1406713	0.0493886	0.07682013

Table 5.2: Diagnostics: Correlation Comparison Table on American Time Usage

Table 5.1 and Table 5.2 tell us the correlations. `Airport Delay` data is highly correlated within its variables, so it make sense that Regression and Stochastic Regression produces the best results. Also, Multiple Imputation gives good results as well. On `American Time Usage` data, variables are really weakly linked with each other, and none of the methods gives results that are highly correlated with the original values.

5.2.2 Difference in Means

Missing %	EM Algorithm	Reg. Imp.	Stoc. Reg. Imp.	MI
5%	-2.506919	1.316287e-13	1.25353e-13	-2.024408e-08
10%	5.139848	1.445702e-13	1.314197e-13	-5.79557e-08
15%	1.235322	2.189939e-13	2.311682e-13	9.21711e-09
20%	-2.760076	-2.700308e-13	-2.991741e-13	3.114098e-07
25%	0.4712096	3.481647e-13	3.393063e-13	-1.657327e-06
30%	0.7751511	1.080179e-13	1.047907e-13	6.817082e-06
35%	1.200881	1.111724e-13	1.174704e-13	-2.153751e-06
40%	-0.767384	-1.707268e-14	-2.252534e-14	0.0003337034
45%	-0.167386	-1.666008e-13	-1.651295e-13	3.803268e-06
50%	0.3401346	-1.075224e-13	-1.064358e-13	0.001762821

Table 5.3: Diagnostics: Difference between real and predicted means Table on Airport data

Missing %	EM Algorithm	Reg. Imp.	Stoc. Reg. Imp.	MI
5%	-2.071277	-0.07625417	-0.6260302	-0.3202157
10%	0.06934133	-0.8806561	-1.390025	-0.9390251
15%	0.2573548	0.3930212	-0.4811219	0.8786178
20%	-0.3660367	-0.5873418	-0.3978093	-0.2552536
25%	0.08603187	0.3734363	0.7578595	0.4061563
30%	-0.2840925	0.1491674	0.9958896	0.8585399
35%	0.3038039	0.1881266	0.06742983	0.461763
40%	-1.270921	-1.158882	-1.443668	-1.298474
45%	0.8032534	1.201826	1.485462	1.178062
50%	-0.3069865	-0.1771209	-0.5442628	-0.672805

Table 5.4: Diagnostics: Difference between real and predicted means Table on American Time Usage data

Table 5.3 and Table 5.4 coincide with the correlation tables, EM give less accurate numbers, and Regression Imputation MI gives better results on the `Airport Delay` data, and the results are less ideal for `American Time Usage` data.

5.2.3 Confidence Intervals Vs. Real Mean

Since the imputation methods give us estimates of the missing data, we can infer the method works well if the true mean is included in the 95% confidence interval of the estimates. A 95% confidence interval can be obtained by the following,

$$\begin{aligned} 0.95 &= P(-z \leq Z \leq z) \\ &= P(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) \\ &= P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) \end{aligned}$$

The results are,

Method	Real Mean	Left Bound	Right Bound
EM Alg. 5% Missing	48.56747	3.2125120	93.92244
EM Alg. 10% Missing	52.36505	9.1407040	95.58940
EM Alg. 15% Missing	50.29527	4.6737314	95.91681
EM Alg. 20% Missing	49.82455	0.1644687	99.48462
EM Alg. 25% Missing	50.70747	3.0884878	98.32645
EM Alg. 30% Missing	51.13314	3.0858718	99.18041
EM Alg. 35% Missing	50.80287	3.3073341	98.29840
EM Alg. 40% Missing	49.58168	1.3714760	97.79188
EM Alg. 45% Missing	51.05148	1.8161559	100.28680
EM Alg. 50% Missing	50.33299	2.1739078	98.49207

Table 5.5: Diagnostics: 95% Confidence Interval for imputed values for EM on Airport Data

Method	Real Mean	Left Bound	Right Bound
Reg. Imp. 5% Missing	48.56747	5.5363562	91.59859
Reg. Imp. 10% Missing	52.36505	5.2120763	99.51803
Reg. Imp. 15% Missing	50.29527	2.9939789	97.59656
Reg. Imp. 20% Missing	49.82455	2.4268980	97.22219
Reg. Imp. 25% Missing	50.70747	2.5906431	98.82429
Reg. Imp. 30% Missing	51.13314	2.5349307	99.73135
Reg. Imp. 35% Missing	50.80287	2.3318270	99.27390
Reg. Imp. 40% Missing	49.58168	1.8599996	97.30335
Reg. Imp. 45% Missing	51.05148	1.8760300	100.22693
Reg. Imp. 50% Missing	50.33299	1.8447759	98.82120

Table 5.6: Diagnostics: 95% Confidence Interval for imputed values for Regression Imputation on Airport Data

Method	Real Mean	Left Bound	Right Bound
Stoc. Reg. Imp. 5% Missing	48.56747	5.5363562	91.59859
Stoc. Reg. Imp. 10% Missing	52.36505	5.2120763	99.51803
Stoc. Reg. Imp. 15% Missing	50.29527	2.9939789	97.59656
Stoc. Reg. Imp. 20% Missing	49.82455	2.4268980	97.22219
Stoc. Reg. Imp. 25% Missing	50.70747	2.5906431	98.82429
Stoc. Reg. Imp. 30% Missing	51.13314	2.5349307	99.73135
Stoc. Reg. Imp. 35% Missing	50.80287	2.3318270	99.27390
Stoc. Reg. Imp. 40% Missing	49.58168	1.8599996	97.30335
Stoc. Reg. Imp. 45% Missing	51.05148	1.8760300	100.22693
Stoc. Reg. Imp. 50% Missing	50.33299	1.8447759	98.82120

Table 5.7: Diagnostics: 95% Confidence Interval for imputed values for Sto. Regression Imputation on Airport Data

Method	Real Mean	Left Bound	Right Bound
MI 5% Missing	48.56747	5.5363562	91.59859
MI 10% Missing	52.36505	5.2120763	99.51803
MI 15% Missing	50.29527	2.9939789	97.59656
MI 20% Missing	49.82455	2.4268983	97.22219
MI 25% Missing	50.70747	2.5906415	98.82430
MI 30% Missing	51.13314	2.5349348	99.73135
MI 35% Missing	50.80287	2.3318205	99.27391
MI 40% Missing	49.58168	1.8603229	97.30303
MI 45% Missing	51.05148	1.8760039	100.22695
MI 50% Missing	50.33299	1.8464670	98.81951

Table 5.8: Diagnostics: 95% Confidence Interval for imputed values for MI on Airport Data

Method	Real Mean	Left Bound	Right Bound
EM Alg. 5% Missing	37.44444	0.65001791	74.23887
EM Alg. 10% Missing	37.15862	1.73526048	72.58198
EM Alg. 15% Missing	38.03687	1.60029621	74.47344
EM Alg. 20% Missing	37.08276	0.87614391	73.28937
EM Alg. 25% Missing	38.05249	1.16436827	74.94060
EM Alg. 30% Missing	37.72414	0.69702363	74.75125
EM Alg. 35% Missing	37.75345	1.21990489	74.28700
EM Alg. 40% Missing	36.92586	-0.47007324	74.32180
EM Alg. 45% Missing	38.28374	1.58422911	74.98326
EM Alg. 50% Missing	37.49379	0.47310702	74.51448

Table 5.9: Diagnostics: 95% Confidence Interval for imputed values for EM on American Time Usage Data

Method	Real Mean	Left Bound	Right Bound
Reg. Imp. 5% Missing	37.44444	0.43035885	74.45853
Reg. Imp. 10% Missing	37.15862	-0.47555249	74.79279
Reg. Imp. 15% Missing	38.03687	0.71795154	75.35578
Reg. Imp. 20% Missing	37.08276	-0.33956418	74.50508
Reg. Imp. 25% Missing	38.05249	0.62918848	75.47578
Reg. Imp. 30% Missing	37.72414	0.35354487	75.09473
Reg. Imp. 35% Missing	37.75345	0.39888277	75.10802
Reg. Imp. 40% Missing	36.92586	-0.96241286	74.81414
Reg. Imp. 45% Missing	38.28374	1.39804011	75.16944
Reg. Imp. 50% Missing	37.49379	0.04227958	74.94531

Table 5.10: Diagnostics: 95% Confidence Interval for imputed values for Regression Imputations on American Time Usage Data

Method	Real Mean	Left Bound	Right Bound
Stoc. Reg. Imp. 5% Missing	37.44444	1.74003642	73.14885
Stoc. Reg. Imp. 10% Missing	37.15862	0.22851812	74.08872
Stoc. Reg. Imp. 15% Missing	38.03687	0.84936301	75.22437
Stoc. Reg. Imp. 20% Missing	37.08276	0.73301243	73.43250
Stoc. Reg. Imp. 25% Missing	38.05249	1.82291885	74.28205
Stoc. Reg. Imp. 30% Missing	37.72414	1.97460611	73.47367
Stoc. Reg. Imp. 35% Missing	37.75345	0.93056210	74.57634
Stoc. Reg. Imp. 40% Missing	36.92586	-0.62442348	74.47615
Stoc. Reg. Imp. 45% Missing	38.28374	2.31821658	74.24927
Stoc. Reg. Imp. 50% Missing	37.49379	0.23903308	74.74855

Table 5.11: Diagnostics: 95% Confidence Interval for imputed values for Stoc. Regression Imputation on American Time Usage Data

Method	Real Mean	Left Bound	Right Bound
MI 5% Missing	37.44444	2.08458767	72.80430
MI 10% Missing	37.15862	0.99355284	73.32369
MI 15% Missing	38.03687	2.23747150	73.83626
MI 20% Missing	37.08276	0.96724547	73.19827
MI 25% Missing	38.05249	1.50571210	74.59926
MI 30% Missing	37.72414	1.89638092	73.55189
MI 35% Missing	37.75345	1.35587494	74.15103
MI 40% Missing	36.92586	-0.44540679	74.29713
MI 45% Missing	38.28374	2.03880267	74.52868
MI 50% Missing	37.49379	0.10766609	74.87992

Table 5.12: Diagnostics: 95% Confidence Interval for imputed values for MI on American Time Usage Data

These tables give us a metric to see wether or not the method is biased. All real mean falls into the 95% confidence interval so the results all three methods produced are unbiased.

CHAPTER 6

Discussions and Conclusions

The objective of this paper is to test which method works better on different kind of dataset with different percentage in missingness. Generally, all of these methods work better on `Airport Delay` data, in which the variable with missing values are highly correlated with combinations of other variables. It is easier to work with missing data with high correlations between its variables.

If researchers encounter missing value problems and they cannot completely ignore those values, the first thing I would recommend is to take a look at the missing value plot included in **MI** package, which can be easily found in R, then look at the internal correlations of its variables. If the data tends to have high correlations among its variables and only a few variables contains missing values, then use any regression imputation method. If there are a large number of variables with missing values, then it looks like Multiple Imputations is the best approach to take. For data with weaker internal correlations,none of these four methods give a good answer, but in comparison EM algorithm gives better results than others. Also, Data with high internal correlations are less sensitive to percentage of data points missing when we use Regression Imputation methods.

Missing value imputations is a very broad topic, this thesis did not test out all possible cases, and it mainly focused on investigating the percentage of missing in data. In the future, people can run analyses on missing values occurring at different blocks, e.g. data missing on a whole ethnic group then draw inferences from there.

Another thing is I only had two datasets for my analysis and they do not cover all types of data available. In the future, people can use datasets coming from various background and use similar approaches to those this thesis took and conclude it with more generalized rules for other researchers to use when they are dealing with missing value problems.

APPENDIX A

Appendix A. Data Explanation

According to the description on data.gov, the airport delay data has the following description, "This table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi-out and taxi-in times, air time, and non-stop distance."

The description on the time usage data on the same source are, "The American Time Use Survey (ATUS) measures the amount of time people spend doing various activities, such as paid work, childcare, volunteering, and socializing."

Bibliography

[1] Andrew Gelman and Jennifer Hill. *Data Analysis, Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013, 2006.

[2] Ph.D. Jeffrey C. Wayman. Multiple imputation for missing data: What is it and how can i use it? *Conference presentation*, 2003.

[3] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 2002.

[4] Yu-Sung Sun, Andrew Gelman, Jennifer Hill, and Masanao Yajima. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistics Software*, 45, 2011.