STAT 854 final project

Abstract

The missing data problem is a common issue in the real world that has attracted attention from the researchers and industrial scientists. A good strategy for handling missing data can imporve the credibility of the data and prediction accuracy. In order to understand the strength of different methods for handling the missing data, in this project, I explored the couple popular approaches and examed their performance using the estimated target mean.

1. Introduction

In many real world applications, the collected data will generally contain some missing value in them. The typical reasons, just to name a few, are due to the mode of data collection, heavy response burden or some privacy concerns. In order to draw precise inference and accurate prediction, people apply different approaches for overcoming the missing data issues.

This project set up a missing data problem that only one study variable (y variable) contains some missing data (None response data). And I addressed the problem by introducing couple of different methods including complete-case analysis, inverse probability weighting, random hot deck and some single imputation methods. The estimated mean of study variable was used for comparing the method performances.

The mechanism of the missing data were discussed in section 2. Then, detailed introduction of each method used in this project was listed in the section 3 methodologies. The section 4 described a simulation study that each method were applied and compared. And section 5 is a conclusion of the project. Acknowledgement and references were listed in section 6 and 7.

2. Missing data mechanisms

Define the response indicators to be

$$R_i = \begin{cases} 1 & \text{, if y\_i is observed} \\ 0 & \text{, if y\_i is missing} \end{cases}$$

So that the probability of responding of i-th observation is

$$p_i = P(R_i | y_i, x_{ij},\ _{j=1,\dots,p})$$

where p is the number of the covariates.

Then three common missing mechanism can be summarized as:

1. Missing completely at random (MCAR)

Study values y's are missed completely at random with a probability $p_i = p$ for all $i \in S$, where S is the total sample set.

2. Missing at random (MAR)

The response probability is associated with the value of the covariates x_{ij} only. So that $p_i = p(x_i)$.

3. Not missing at random (NMAR)

In this mechanism the missing probability of study values are associated with both x_i and y_i. This means that $p_i = p(x_i, y_i)$.

\

This project focused on mechnism MCAR and MAR. And performance of each method introduced in section 3 was tested and compared accordingly in section 4.

3. Methodologies

3.1 Complete-case analysis

A naive approach is to use complete-case analysis method which analysis the data basing only on the observations that has no missing values (Response data, S_R).

3.2 Inverse probability weighting (IPW)

The IPW method still uses all the observations from the group of response data, S_R and weights them with the estimated

$$\hat{p}_i = P(R_i = 1 | y_i, x_{ij},\ _{j=1,\dots,p})$$

Then the estimation of the mean of the study variable μ_y will be

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{i \in S_R} \frac{y_i}{p_i}$$

This approach is similar to the Horvitz-Thompson estimator. And $\hat{p}_i$ can be estimated using logistic regression on the responsed data, S_R.

3.3 Single imputation methods

Single imputations refers to imputation methods that replace the missing value y_i with one imputed values y_i^*. These type of methods require imputation models but

the choice of the model depends on how much auxiliary information people can get from the data. Some typical methods were used to exam their performance in this project:

3.3.1 Mean imputation

Define the total dataset to be consisted of response data and non-response data as:

$$S = S_R \cup S_N$$

When there is no other auxiliary information is available, one can use imputate each missing value $y_j, j \in S_M$ by $\overline{y}_R$ as:

$$y_j^* = \overline{y}_R = \frac{1}{n_R} \sum_{i \in S_R} y_i \,,\ j \in S_M$$

3.3.2 Random hot deck imputation

For every observation $j \in S_M$, this method impute the missing value y_j by randomly pickup a value from the assumed similar group of observations which is S_R. ie., $y_j^* = y_k$ where k is randomly selected from S_R.

3.3.3 Regression imputation

If the auxiliary information on covariates $x_{ij},\ _{i=1,\dots,n; j=1,\dots,p}$ can be observed for every observation $i \in S$. $y_i, i \in S_R$ are the study values observed and $y_i, i \in S_N$ are missed.

The missing value will be imputed by model $f(X_N, \beta_R)$ where X_N is the covariate matrix of the coresponding nonresponse observations and β_R is the coefficient vector estimated using the response data.

Under the linear assumption, one can choose to use a linear regression model as:

$$Y_N = X_N \hat{\beta}_R + \epsilon$$

where the ϵ is assumed following a Multinormal distribution.

The coefficient vector β_R is the vector that estimated using least square estimation as:

$$\hat{\beta}_R = (X_R{}' X_R)^{-1} X_R{}' Y_R$$

where X_R is the covariate matrix of response data, and Y_R is the column vector of the study value of the response data.

Notice that, when calculating $(X_R{}' X_R)^{-1}$, $X_R{}' X_R$ is a positive semi-definite matrix. For computing concerns, it can be computed efficiently using Cholesky decomposition:

$$\text{Cholesky Decomposition:} \quad X_R{}' X_R = LL^*$$

where L is a lower triangular matrix with real and positive diagonal entries, and L^* denotes the conjugate transpose of L. And in real number case, L^* is the transpose of L, ie. $L^* = L'$.

So that

$$(X_R{}'X_R)^{-1} = (LL')^{-1} = (L^{-1})'L^{-1}$$

where the L^{-1} can be easily solved by a forwardsolve method for matrix solving.

$$\hat{\beta}_R = (X_R{}'X_R)^{-1}X_R{}'Y_R = (L^{-1})'L^{-1}X'_RY_R$$

And the imputed study value for the missing data Y_N can be estimated as:

$$\hat{Y}_N = X_N\hat{\beta}_R = X_N(L^{-1})'L^{-1}X'_RY_R$$

3.3.4 K-nearest neighbor imputation

K-nearest neighbor (KNN) algorithm is a non-parametric model that assumes the similar observations are closer to each other. For an observation that we want to predict, KNN look at the most similar observations or nearest neighbors and then make prediction using their study value. The auxiliary information on the covariates of all of the data are necessary for imputation.

The whole imputation algorithm can be broke down into the following four steps:

- Step 1: For a non-response datum we want to impute, compute the distance of the datum to all of the response data.
- Step 2: Find the nearest K neighbors to the non-response datum.
- Step 3: Make prediction based on selected K neighbors.
- Step 4: Repeat step 1 to 3 for the rest of the non-response datum we want to impute.

The distance metric used in this project is the L1-norm which is the absolute difference of the components of covariate vectors.

The prediction is made by taking the mean value of the K neighbors.

The number of neighbors, K, is a tuning parameter that need to be tuned using cross-validation. In this project, I tuned the value of K using a Leave-One-Out (LOO) cross-validation and minimizing the L1-loss on the response data.

The Leave-One-Out algorithm used in the project for find K can be summarized as follows:

- Step 1: Leave out one observation $j \in S_R$ from the response data.
- Step 2: Fit the KNN model on the remaining data and predict $\hat{Y}_j$ for different choices of K.
- Step 3: Repeat step 1 and 2 for all the rest of the response data.
- Step 3: Compute the total L1-loss for every choice K across the all the observation.
- Step 4: Choose the K value that associated with the smallest L1-loss.

3.3.6 Imputation classes

This is a method that divide the whole dataset into different groups, and then impute the missing data within each group using one of the method specified above.

If the auxiliary information on x is available for all $i \in S$. Define the whole dataset to be divided by k subset as:

$$S = S_1 \cup S_2 \cup \ldots \cup S_k$$

where $S_i \cap S_j = \phi,\ i, j \in 1, \ldots, \mathrm{k}$

Each subset S_i which is the imputation class, consists of a group of respond data and a group of non-response data as:

$$S_i = S_{iR} \cup S_{iN},\ i \in 1, \ldots, \mathrm{k}$$

And the further imputation using the methods mentioned above can be used within each S_k as before. In this project, I combined 12 imputation classes with mean imputation, random hot deck imputation, linear regression imputation and KNN imputation together.

4. Simulation

4.1 Setups

All the methods mentioned above were used in this section for testing their performance under both of the missing data mechnisms, MCAR and MAR.

The setups of the simulated data set were:

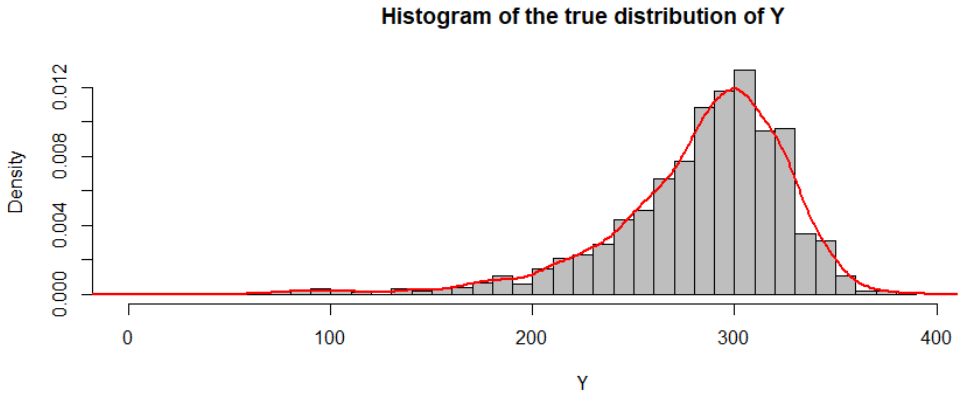
(1) The total number of observations was $N = 1000$.

(2) There were 3 covariates: x1, x2, x3. x1 followed an exponential distribution of Exp(1). x2 followed a binomial distribution of Binom(5,0.7). x3 followed a normal distribution of Normal(30,5).

(3) Each observation y_i was constructed with the relationship: $y_i = 300 + 6x_{i1} + 10x_{i2} + 3x_{i2} - 1.5x_{i1}x_{i3} - x_{i2}x_{i3} + \epsilon_i$ where ϵ_i followed a normal distribution of Normal(0,1).

The interaction terms were constructed to simulate the practical situations where covariates were interacted with each other. And covariates followed different kinds of distribution to represent the values collected from different type of questions in the survey.

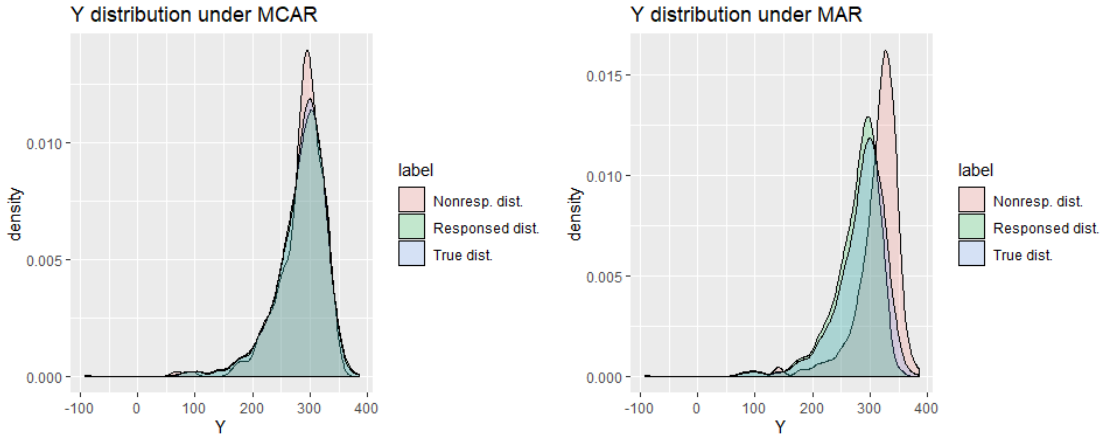
The distribution of the study variable was visualized in the following figure. It is right skewed with a long tail on the left.



(4) The proportion of the data missing from the population was 20%.

(5) Under MCAR the probability of response was $p = 0.8$. Each observation had an probability of 0.2 being missing.

(6) Under MAR the probability of response was $p_i = 1 - 1/(1 + exp(+3x_{i3} + 6x_{i2}x_{i3}))$. Whether the study values were responsed depend on x_3 and the interatction of x_2 and x_3.



(7) The imputation models, which could include auxiliary information, only used covariates x_1,x_2 and x_3 without any interaction. This simulated the situations where people didn't know the true relationship between the data covariates.

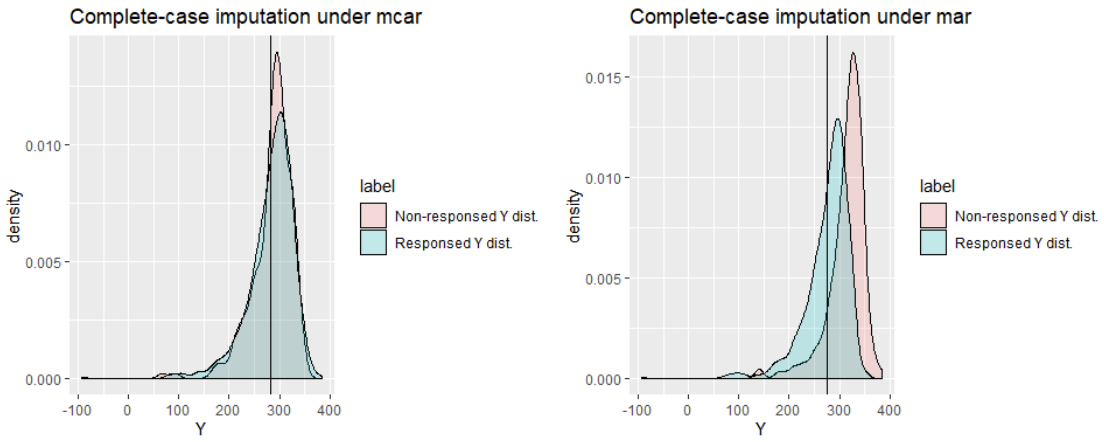
4.2 Test model performance

In the following sections, the distribution of the estimated study variable, distribution of the only responsed study variable and distribution of the only non-responsed study variable were visualized for both MCAR and MAR situation.

Then the estimated means of the study variable and the absolute differences to the ture mean under both MCAR and MAR mechanism were summarized in the table.

4.2.1 Complete-case analysis

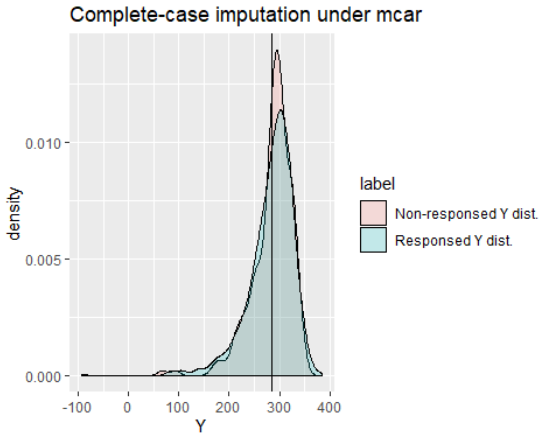
The estimated mean of the study values of the whole data set were visualized by the vertical lines below.

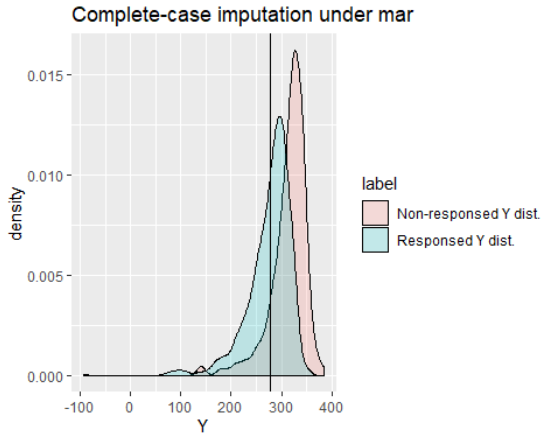


Measure	Under MCAR	Under MAR
Estimated Mean	282.606	275.664
abs(est. - true)	0.523	7.465

4.2.2 Inverse probability weighting

The estimated mean of the study values of the whole data set were visualized by the vertical lines below.

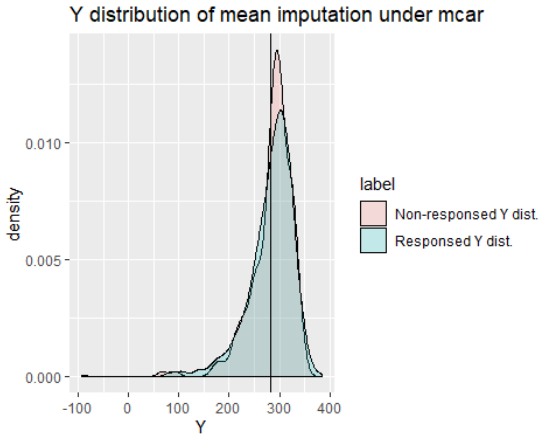


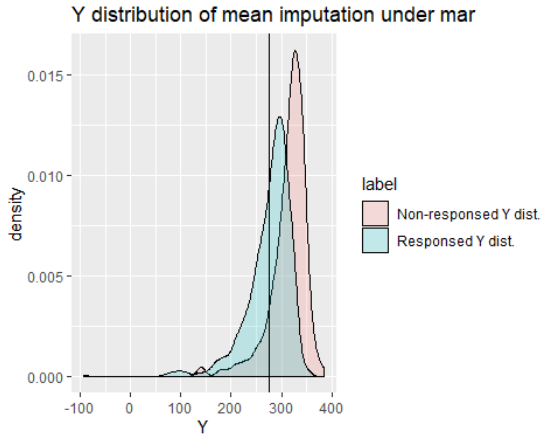


Measure	Under MCAR	Under MAR
Estimated Mean	284.889	278.422
abs(est. - true)	1.759	4.707

4.2.3 Mean imputation

The estimated mean of the study values of the whole data set were visualized by the vertical lines below.

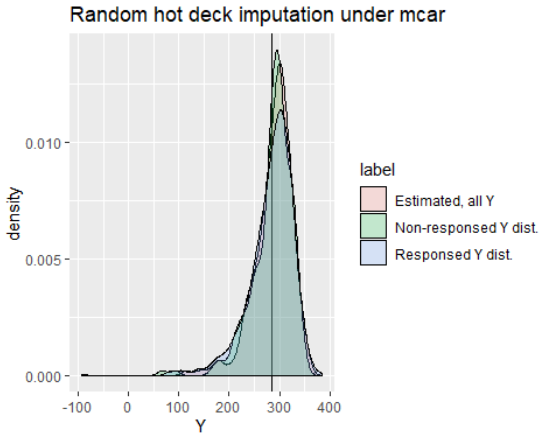


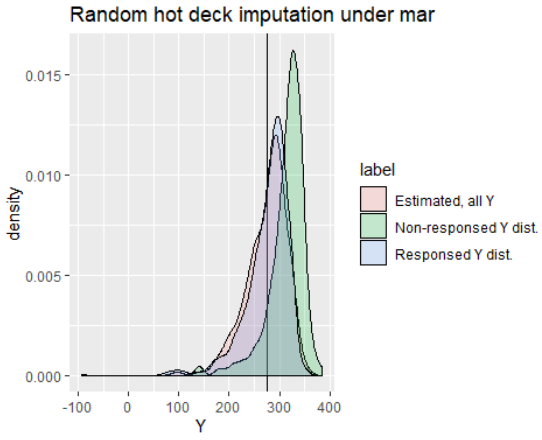


Measure	Under MCAR	Under MAR
Estimated Mean	282.605	275.664
abs(est. - true)	0.522	7.465

4.2.4 Random hot deck imputation

The estimated mean of the study values of the whole data set were visualized by the vertical lines below.

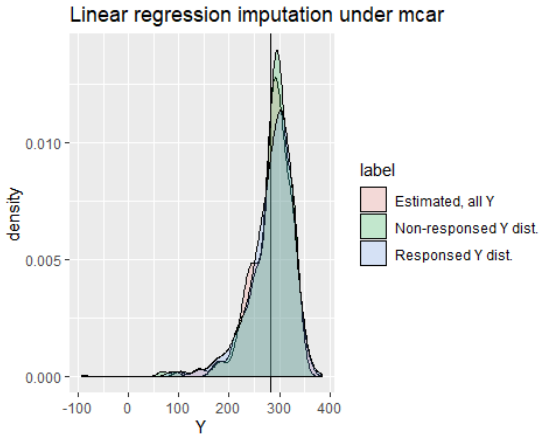


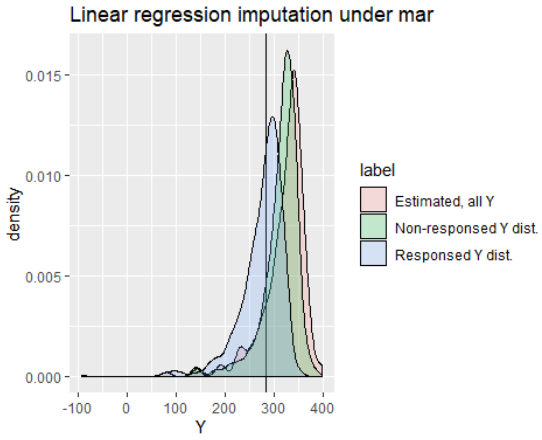


Measure	Under MCAR	Under MAR
Estimated Mean	281.152	275.870
abs(est. - true)	1.977	7.259

4.2.5 Linear regression imputation

The estimated mean of the study values of the whole data set were visualized by the vertical lines below.

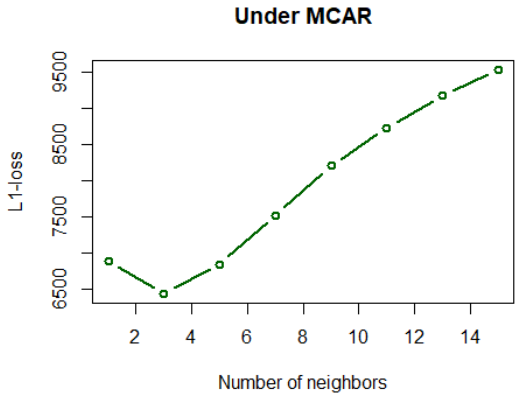


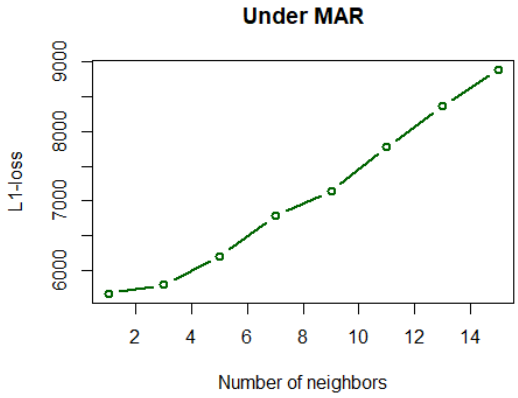


Measure	Under MCAR	Under MAR
Estimated Mean	282.920	284.470
abs(est. - true)	0.209	1.341

4.2.6 KNN imputation

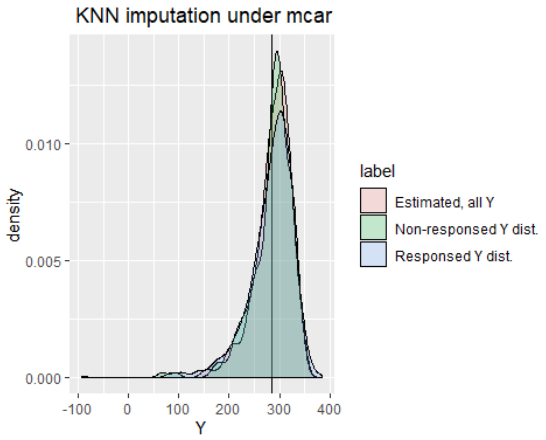
The optimal number of neighbors, K, was obtained by finding the minimum L1-norm, and the relationships of the K and the L1-loss were visualized below:

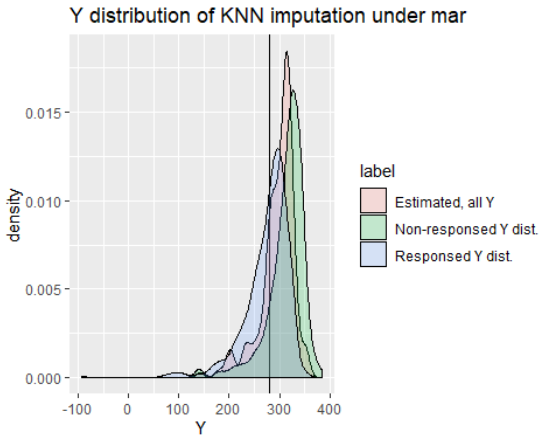




Under MCAR, the optimal number of neighbors is $K = 3$. Under MAR, the optimal number of neighbors is $K = 1$.

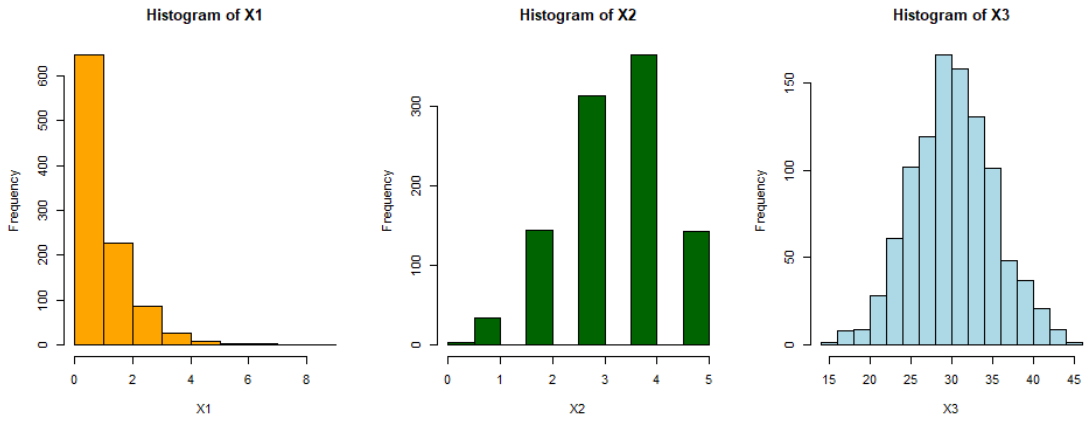
The estimated mean of the study values of the whole data set were visualized by the vertical lines below.





Measure	Under MCAR	Under MAR
Estimated Mean	283.561	279.823
abs(est. - true)	0.432	3.306

4.2.7 Imputation classes



The distribution of three explanatory variables were visualized above. Basing on that,

- Variable x_1 was divided to 2 groups depending on wether $x_1 \leq 1.5$.
- Variable x_2 was divided to 2 groups depending on wether $x_2 \leq 3$.
- Variable x_3 was divided to 3 groups depending on wether $x_3 \leq 25$, $25 < x_3 \leq 35$ and $x_3 > 35$.

Then four imputation methods, the mean imputation, random hot deck imputation, linear regression imputation and KNN imputation, were combined seperately with imputation classes for imputing the missing values.

In the end, the results were summaried using tables below.

Under MCAR,

Measure Under MCAR	Mean imputation	Random hot deck	Linear regression	KNN
Estimated Mean	283.218	282.239	283.061	283.973
abs(est. - true)	0.0889	0.8901	0.068	0.8439

\

Under MAR,

Measure Under MAR	Mean imputation	Random hot deck	Linear regression	KNN
Estimated Mean	314.469	313.459	313.312	315.622
abs(est. - true)	31.340	30.330	30.183	32.492

5 Conclusion

As we can see, under the MCAR mechanism, no matter which method we use, the estimated mean of the study variable is quite similar to the true mean.

Especially, the linear regression imputation combined with imputation class imputed the study variable that generated a closest mean value to the true mean. Mean imputation generated a similar result as random hot deck method's. Moreover, the imputation class can reduce the absolute bias of the estimated mean for every single method.

Under MAR, the performance of each method didn't seem to be as good as that under MCAR.

The linear regression imputation still imputed the missing data with lowest absolute bias. But the imputation class method enlarged the bias for every single model. The potential reason could be that by constructing the classes, the complex interaction between the covariate were ignored. Thus, worse results were generated.

The remark from this project is that under the framework of the single imputation, one can choose to use linear regression imputation if the auxiliary information is available. If no auxiliary information available, then random hot deck or mean imputation are both good choice for handling the missing data. And the imputation class need to be used with cautious. Because if the data contains interactions between the variables, the method may not impute the missing data well.

6 Acknowledgement

The equations and methods mentioned in this project are partially from the in Prof. Changban's lecture notes. Thank you Prof Changbao for teaching a great course.

7 References

Hastie T., Tibshirani R., Friedman J. (2008). The Elements of Statistical Learning (Second Edition). Springer Series in Statistics.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Frederick Novomestky (2012). matrixcalc: Collection of functions for matrix calculations. R package version 1.0-3. https://CRAN.R-project.org/package=matrixcalc