IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 18, Issue 2 Ser. I (Mar. – Apr. 2022), PP 52-62
www.iosrjournals.org

Handling Missing Values in Dataset

Bibi Sherriza Ali (MSc , BSc)
Department of Mathematics, Physics, and Statistics, University of Guyana (Berbice Campus)

Abstract
This paper analyses method for handling missing data in research. All researchers have faced the problem of missing quantitative data at some point in their work. Research informants may decline or forget to answer a survey question, as such files are lost, or data are not recorded properly. Given the expenditure of collecting data, we cannot afford to start over or wait until we have developed perfect methods of assembling information. We find ourselves left with the decision of how to analyze data when we do not have complete information from all informants. Researchers either intentionally or by default in a statistical analysis drop informants who do not complete data on the variables of interest. As an alternative to complete-case analysis, researchers may find a plausible value for the missing observations, such as using the mean of the observed cases on that variable.
I will argue that all researchers need to be cautioned when faced with missing data. Methods for analyzing missing data require assumptions about the nature of the data and about the reasons for the missing observations that are often not acknowledged. When researchers use missing data methods without carefully considering the assumptions required of that method, they run the risk of obtaining biased and misleading results. Reviewing the stages of data collection, data preparation, data analysis, and interpretation of results will highlight the issues that researchers must consider in making a decision about how to handle missing data in their work. This paper focuses on commonly used missing data methods: exclusion, simple imputation, and model-based imputation.

Keywords: *Exclusion, Simple Imputation, Model-Based imputation, Quantitative data, complete case analysis*

Date of Submission: 02-03-2022 Date of Acceptance: 13-03-2022

I. Introduction

In almost any research you perform, there is the potential for missing or incomplete data. Missing data can occur for many reasons: participants can fail to respond to questions (legitimately or illegitimately), subjects can withdraw from studies before they are completed, and data entry errors can occur. The issue with missingness is that nearly all classic and modern statistical techniques assume (or require) complete data and most common statistical packages default to the least desirable options for dealing with missing data: deletion of the case from the analysis. Most people analyzing quantitative data allow the software to default to eliminating important data from their analyses, despite that individual or case potentially having a good deal of other data to contribute to the overall analysis. The most common way of handling missing data is called-wise deletion i.e. delete case (or rows/list) containing missing values and running a model, and using data set without missing values (known as the complete case analysis. What is wrong with deletion? The problems are twofold :(1) loss of information (i.e. reduction in statistical power) and (2) potential bias in parameter estimates under most circumstances. Whether a data user is an

Experienced statisticians or business analysts, he or she must be able to assess the prevalence of missing data and to identify appropriate methods to address it.

i. Dataset Component

Two datasets were used in this research: Dataset – Votes and Dataset – Marketing. The following table is showing the component of these datasets.

Dataset: Votes.repub	Dataset: Marketing
Number of non – missing values = 1333	Number of non – missing values = 123208
Number of missing values = 217	Number of missing values 2694
Proportional missing values = 0.14	Proportional missing value = 0.02

Table 1.0: Showing dataset component

When looking at the proportion of missing values in comparison with the portion of missing observations there is a great difference, the question one needs to ask when handling missing data, should we just delete the missing observations or should one replace them with central tendency.

Deleting the observation might not be a good method since doing this can leave us with no data to work with.

A missing observation is considered a row with a least one entry missing, which may indicate a large portion of missingness even if there isn't, whereas missing values would account for every missing entry. This, therefore, means that missing values will give a true portion of missingness in a data set. I would conclude that missing values is more revealing than missing observation.

Table 2: Central Submatrix for dataset: Marketing. Here the first 5 observations of the submatrix for the dataset marketing is provided.

	Age	Edu	Occupation	Lived	Dual_Income	Household	Householdu18
2248	2	4	6	5	2	2	2
2249	1	2	5	5	1	3	0
2250	4	4	1	4	3	1	0
2251	3	4	5	5	3	3	0
2252	3	4	4	5	1	3	0
2253	4	5	1	4	1	3	0

ii. Correlation of the dataset.

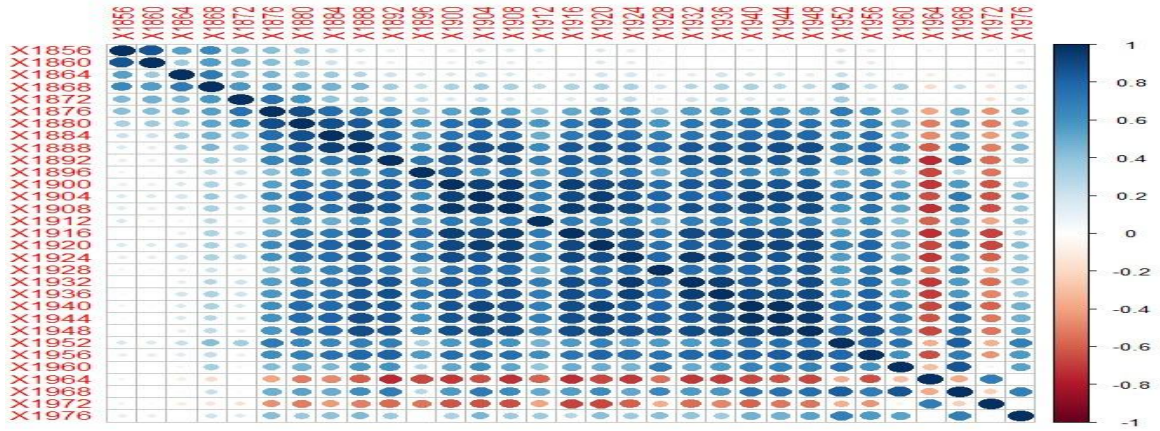

Figure 1.1.Correlation plot for data set: Votes.

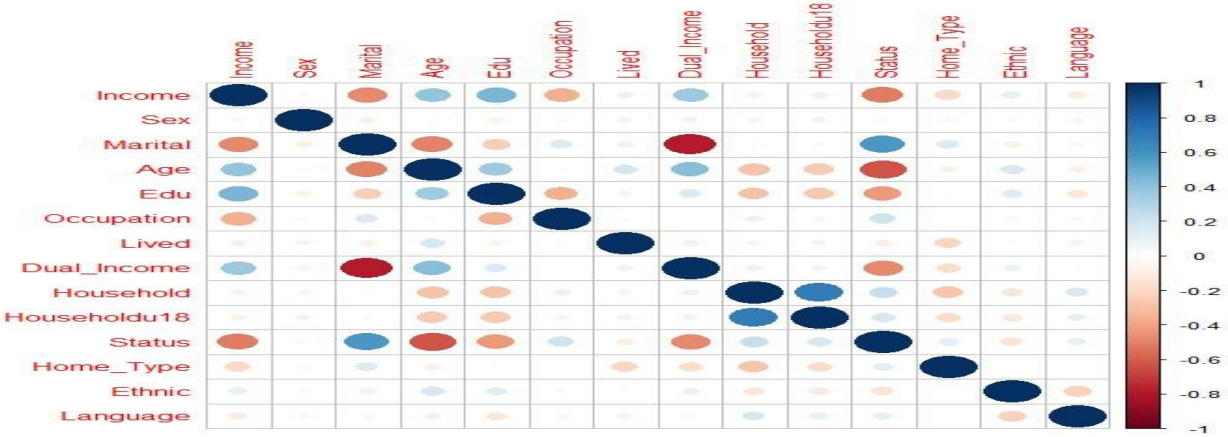

Figure 1.2.Correlation plot for data set: Marketing.

Figure 1.1, shows that most of the variables are strongly correlated in a positive direction, with only few variables that are negatively correlated.

Figure 1.2, indicated a few variables that are strongly correlated in a negative direction and not much variables strongly correlated in the positive direction. It can be concluded that most variables are not highly correlated.

iii. Frobenius norm of the difference matrix.

Frobenius Norm for data set: Votes

$$\|(\square_{\square}) - \square\square\square(\square_{\square})\|_{\square} = 2559.66$$

Frobenius Norm for data set: Marketing

$$\|(\square_{\square}) - \square\square\square(\square_{\square})\|_{\square} = 0.42$$

The Frobenius norm of the difference matrix for the data set vote is very large in comparison with the data set marketing, which implies that there is a significant difference in the covariance of $\square_{\square}$ relative to the covariance of $\square_{\square}$ for the data set vote. While on the other hand there is a small difference in the covariance of $\square_{\square}$ relative to the covariance of $\square_{\square}$ for the data set marketing.

For data set: Votes

$$|(\square^{\square}{}_{\square} \ \square_{\square}) - \square\square\square(\square^{\square}{}_{\square} \ \square_{\square})| = 1.215791e+91$$

For data set: Marketing.

$$|(\square^{\square}{}_{\square} \ \square_{\square}) - \square\square\square(\square^{\square}{}_{\square} \ \square_{\square})| = 5.797325e+57$$

The difference in the determinant for data set votes and marketing is very large, with dataset vote being a larger.

iv. The original data is generated and stored the data matrix X.

```
          [,1]        [,2]       [,3]        [,4]         [,5]       [,6]         [,7]      [,8]        [,9]
1,] -0.637456732  6.52607664 -4.640393  6.217929735 -0.735872773 2.61770147  1.42565500  6.284164  1.30079402
2,] -2.816495336  1.72018111 -5.629618  6.996712953  2.014640419 0.10545222  2.04577507  8.502451  1.40639356
3,] -1.436540312  1.33078129 -5.799672  2.451259223 15.321249063 0.20373406  3.39741184  8.030983 -3.15093392
4,] -0.679821562 -0.55167955 -4.730765  4.251998450  1.796710222 1.05678349 10.88655625  8.407999  1.27167051
5,] -2.376133520  3.00972936 -2.752434  3.248565083  3.131480605 1.58185632  2.44162561  7.599430 -1.19850317
```

Vector of different rate of missingness between 0 and 1

$\epsilon =$

```
[1] 0.00000000 0.02564103 0.05128205 0.07692308 0.10256410 0.12820513 0.15384615 0.17948718 0.20512821
10] 0.23076923 0.25641026 0.28205128 0.30769231 0.33333333 0.35897436 0.38461538 0.41025641 0.43589744
19] 0.46153846 0.48717949 0.51282051 0.53846154 0.56410256 0.58974359 0.61538462 0.64102564 0.66666667
28] 0.69230769 0.71794872 0.74358974 0.76923077 0.79487179 0.82051282 0.84615385 0.87179487 0.89743590
37] 0.92307692 0.94871795 0.97435897 1.00000000
```

v. **Correction Matrix for X_e =**

```
            [,1]        [,2]        [,3]        [,4]         [,5]        [,6]        [,7]        [,8]        [,9]
[1,]  1.00000000 -0.07088264 -0.12365318 -0.08592739  0.096665683 -0.115196903  0.06426101 -0.08105688 -0.13083817
[2,] -0.07088264  1.00000000 -0.21858594 -0.08718510  0.071768480  0.063847955  0.12202306 -0.18714257 -0.25641918
[3,] -0.12365318 -0.21858594  1.00000000  0.13536079 -0.036050721 -0.155978283 -0.08088620  0.17095033  0.10297188
[4,] -0.08592739 -0.08718510  0.13536079  1.00000000 -0.152807563  0.037931331 -0.07185615  0.02495286  0.20043140
[5,]  0.09666568  0.07176848 -0.03605072 -0.15280756  1.000000000 -0.004059491 -0.11877957 -0.22249014 -0.25472565
[6,] -0.11519690  0.06384796 -0.15597828  0.03793133 -0.004059491  1.000000000 -0.04440207 -0.16368114  0.17150362
[7,]  0.06426101  0.12202306 -0.08088620 -0.07185615 -0.118779572 -0.044402067  1.00000000  0.03412602 -0.08419645
[8,] -0.08105688 -0.18714257  0.17095033  0.02495286 -0.222490141 -0.163681136  0.03412602  1.00000000  0.05758969
[9,] -0.13083817 -0.25641918  0.10297188  0.20043140 -0.254725651  0.171503619 -0.08419645  0.05758969  1.00000000
```

Correction Matrix for X_c =

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	-2.474951414	0.484596458	-4.3897481	3.17585072	5.135401558	-0.798634519	7.01981578	9.565546	0.579621797
[2,]	-0.976331914	2.895645036	-4.3119217	4.17159899	5.992716675	-0.764456469	-0.32838405	7.904732	1.741519168
[3,]	-1.845524144	3.847593354	-2.7807643	5.07369723	4.251826981	-0.025002959	3.68948114	8.746527	0.669801838
[4,]	1.009807836	-3.106435330	-3.8685138	6.51927556	9.692846629	0.949386526	4.04344043	8.733293	-0.701553708
[5,]	0.204248610	0.121604959	-3.7591931	4.83930018	3.994548673	0.925071123	4.06265461	8.143998	0.157784450

vi. Superposed plots of the relative loss of correlation structure as a function of the rate of missingness.

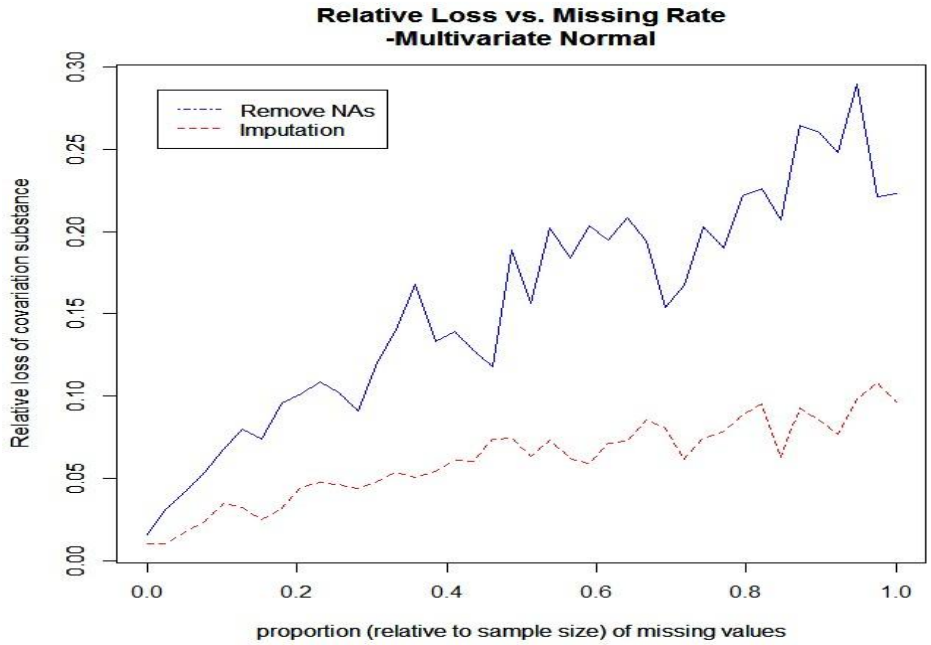

Figure 2.1.Relative Loss vs Missing rate for Multivariate Normal Distribution.

v. Superposed plots of the relative loss of correlation structure using determinant.

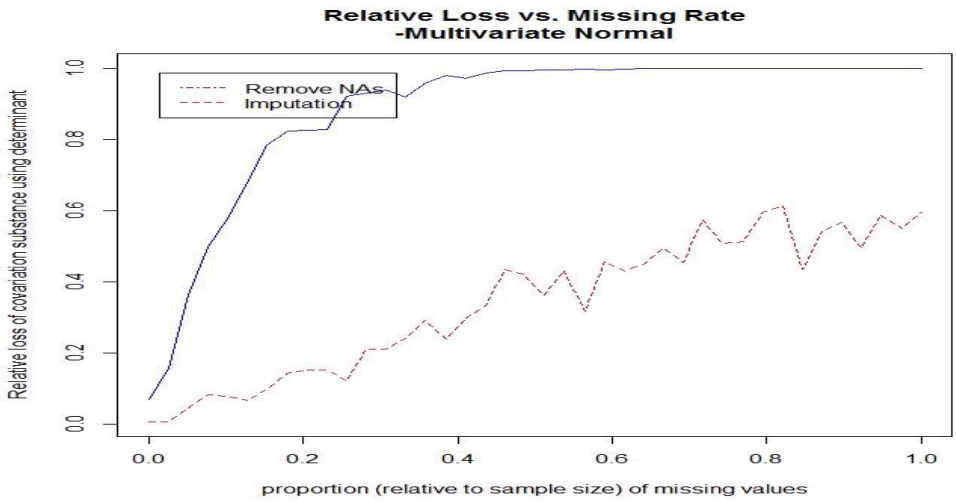

Figure 2.2.Relative Loss vs Missing rate for Multivariate Normal Distribution using Determinant.

From figure 2.1, it is quite evident that when removing the NAs, the relative loss seems to increase significantly when compare to the method of imputing central tendency. When calculating the relative loss using the determinant the method of removing NA's converges to 1 as shown in figure 2.2.

2. **Generating data from a Multivariate Gaussian as in part 1, but this time with a different mean and a different covariance matrix, namely μ = 0 and Use ρ = 0.75.**

Correction Matrix for =

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1.0000000	0.8402686	0.7497326	0.7637620	0.7996424	0.8001264	0.8054731	0.7985003	0.7737268
[2,]	0.8402686	1.0000000	0.7839332	0.7968690	0.8090420	0.7696957	0.7849076	0.7505928	0.7638307
[3,]	0.7497326	0.7839332	1.0000000	0.7193198	0.7991581	0.7056120	0.7465053	0.7223183	0.7463821
[4,]	0.7637620	0.7968690	0.7193198	1.0000000	0.7816508	0.7770778	0.7526895	0.8026664	0.7091619
[5,]	0.7996424	0.8090420	0.7991581	0.7816508	1.0000000	0.7499450	0.8030756	0.8071719	0.7778523
[6,]	0.8001264	0.7696957	0.7056120	0.7770778	0.7499450	1.0000000	0.7460558	0.8010112	0.7555198
[7,]	0.8054731	0.7849076	0.7465053	0.7526895	0.8030756	0.7460558	1.0000000	0.7657298	0.7587680
[8,]	0.7985003	0.7505928	0.7223183	0.8026664	0.8071719	0.8010112	0.7657298	1.0000000	0.7356073
[9,]	0.7737268	0.7638307	0.7463821	0.7091619	0.7778523	0.7555198	0.7587680	0.7356073	1.0000000

Correction Matrix for =

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1.0000000	0.6907206	0.6922963	0.6752497	0.5914961	0.6413622	0.7361672	0.7081212	0.6651417
[2,]	0.6907206	1.0000000	0.6801785	0.6712596	0.6180329	0.6746656	0.6684745	0.6618390	0.6602593
[3,]	0.6922963	0.6801785	1.0000000	0.6628174	0.6567918	0.6778246	0.6909569	0.6828932	0.6718667
[4,]	0.6752497	0.6712596	0.6628174	1.0000000	0.6217648	0.6779845	0.6847635	0.6800992	0.6343552
[5,]	0.5914961	0.6180329	0.6567918	0.6217648	1.0000000	0.6413513	0.6425710	0.6705019	0.6610489
[6,]	0.6413622	0.6746656	0.6778246	0.6779845	0.6413513	1.0000000	0.6479197	0.6764002	0.6309099
[7,]	0.7361672	0.6684745	0.6909569	0.6847635	0.6425710	0.6479197	1.0000000	0.6680117	0.6929748
[8,]	0.7081212	0.6618390	0.6828932	0.6800992	0.6705019	0.6764002	0.6680117	1.0000000	0.6786092
[9,]	0.6651417	0.6602593	0.6718667	0.6343552	0.6610489	0.6309099	0.6929748	0.6786092	1.0000000

i **The superposed plots of the relative loss of correlation structure as a function of the rate of missingness.**

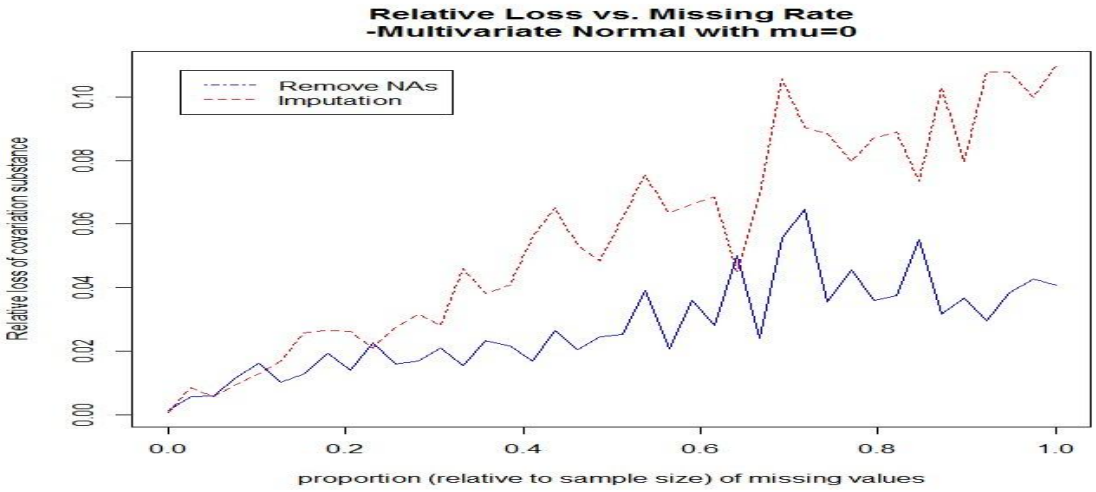

Figure 2.3.Relative Loss vs Missing rate for Multivariate Normal Distribution.

ii. Calculating the relative loss in correlation Structure using the determinant.

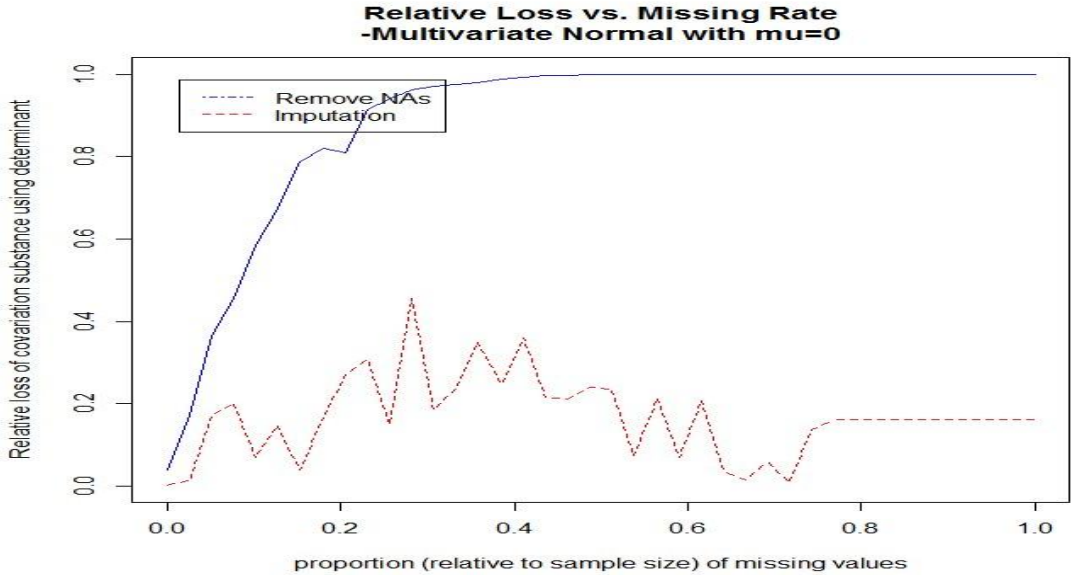

Figure 2.4.Relative Loss vs Missing rate for Multivariate Normal Distribution, mu=0 using Determinant.

From figure 2.3, when dealing with a Multivariate Normal distribution with a mean of zero, it is evident that the relative loss when using the method of central tendency is very high in comparison to the method of removing NA's, removing NA's reaches approximately 5% while the method of central tendency exceeds 10 %.Here we can say that the method of central tendency does not have any effect as shown in figure 2.1, it can be concluded that since □ = □. □□ the variables will be highly correlated making the method of removing NA's not as significant as shown in figure 2.1.

When using the difference in determinant to generate the relative loss, the relative loss for the technique of removing missing rows quickly converges to 1, the relative loss for the technique of central imputation is far from converging to 1.

Based on figure 2.1, when using the Frobenius Norm to generate the relative loss, the technique of removing missing rows relative loss increase significantly in comparison to the technique of imputing central tendency except for figure 2.3, which is because of high correlation among the variables the method of central imputation wouldn't have a significant effect as shown in figure 2.1.

Correction Matrix for □□ =

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1.00000000	0.08112717	0.148216322	0.09874299	0.182492198	0.02772545	-0.092314280	-0.057748157	-0.13853927
[2,]	0.08112717	1.00000000	0.028230113	-0.14910022	0.057046844	-0.13820461	0.141049897	-0.137293620	-0.02277917
[3,]	0.14821632	0.02823011	1.000000000	0.19283299	0.075104579	0.12892965	0.003082179	-0.057771824	0.09677330
[4,]	0.09874299	-0.14910022	0.192832985	1.00000000	0.223043951	-0.01454630	0.093535447	0.017793001	0.19302178
[5,]	0.18249220	0.05704684	0.075104579	0.22304395	1.000000000	-0.09455943	0.002504564	0.139228094	-0.02691168
[6,]	0.02772545	-0.13820461	0.128929653	-0.01454630	-0.094559426	1.00000000	-0.139021247	-0.060946524	-0.09739710
[7,]	-0.09231428	0.14104990	0.003082179	0.09353545	0.002504564	-0.13902125	1.000000000	-0.003717091	0.12744755
[8,]	-0.05774816	-0.13729362	-0.057771824	0.01779300	0.139228094	-0.06094652	-0.003717091	1.000000000	0.03778798
[9,]	-0.13853927	-0.02277917	0.096773302	0.19302178	-0.026911680	-0.09739710	0.127447547	0.037787978	1.00000000

Correction Matrix for □□ =

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1.000000000	0.03497915	0.095574379	-0.026025976	0.06767409	-0.030504153	-0.084707076	-0.007407746	0.037445069
[2,]	0.034979145	1.00000000	-0.017934550	-0.043668899	-0.03863724	0.067389668	-0.013557704	-0.093097034	-0.079091570
[3,]	0.095574379	-0.01793455	1.000000000	0.162257484	-0.04168117	-0.019362488	0.026122797	0.025563673	0.006771285
[4,]	-0.026025976	-0.04366890	0.162257484	1.000000000	0.01765165	-0.002852103	0.004467986	-0.075922297	0.036166201
[5,]	0.067674092	-0.03863724	-0.041681173	0.017651654	1.00000000	-0.022509197	-0.102587723	0.144742169	0.028261770
[6,]	-0.030504153	0.06738967	-0.019362488	-0.002852103	-0.02250920	1.000000000	-0.078417256	-0.011536252	-0.009018785
[7,]	-0.084707076	-0.01355770	0.026122797	0.004467986	-0.10258772	-0.078417256	1.000000000	0.003505570	0.078086840
[8,]	-0.007407746	-0.09309703	0.025563673	-0.075922297	0.14474217	-0.011536252	0.003505570	1.000000000	0.092182653
[9,]	0.037445069	-0.07909157	0.006771285	0.036166201	0.02826177	-0.009018785	0.078086840	0.092182653	1.000000000

iii. The superposed plots of the relative loss of correlation structure as a function of the rate of missingness.

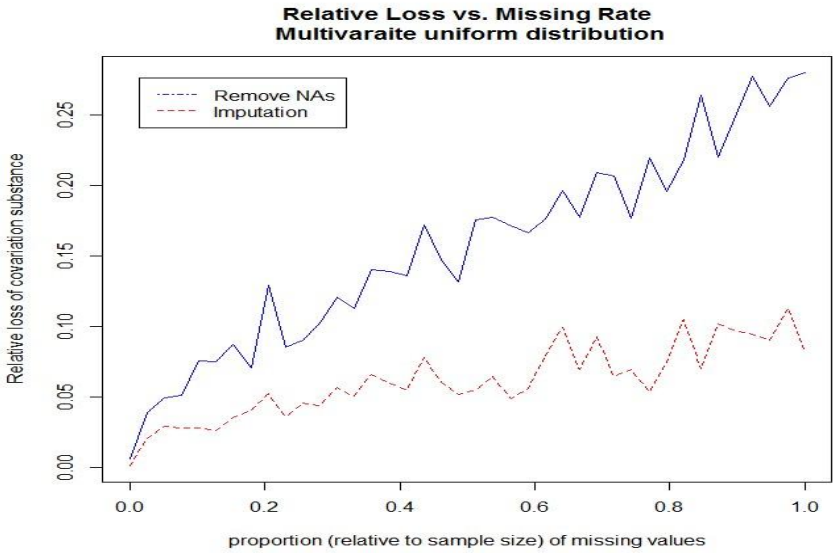

Figure 2.5.Relative Loss vs Missing rate for Multivariate uniform Distribution.

iv Calculating the relative loss in correlation structure using the determinant.

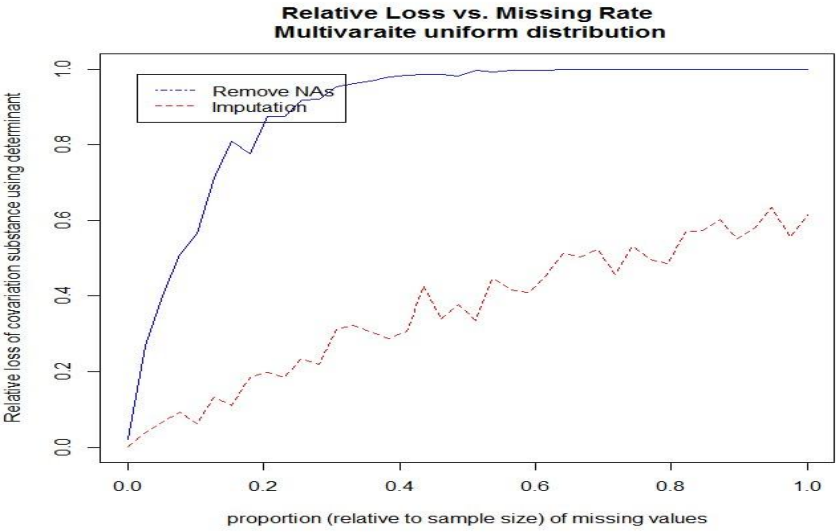

Figure 2.6.Relative Loss vs Missing rate for Multivariate uniform Distribution, using determinant.

From the multivariate uniform distribution, when removing missing values from a data set there is a significant relative loss in comparison with imputing values(like median, mode or mean).From figure 2.5,we can see that there is a large gap between relative loss and proportion(relative to sample size)of missing values among these two methods. Again when using the determinant the method of deleting missing values converges to 1.

The change in the generating distribution between the multivariate normal (Gaussian) distribution and multivariate uniform distribution seems to have minimal change in performance of the loss functions. The multivariate normal (Gaussian) distribution and multivariate uniform distribution shown in figure 2.1 and 2.3 tends to produce the same result when using Frobenius Norm, this is also true when using the difference in determinants.

Imputing measures of central tendency (mean, median and mode) for missing values will result in less relative loss in comparison of removing rows of missing values. When using the Frobenius Norm, the method of removing NA's result in a significant relative loss in comparison to imputing central tendency but there seems to be an adverse effect when there is a high correlation among the variables.

When using the differences in the determinants method to analyze relative loss, the relative loss for the method of removing NA's converges quickly to 100%.

3.) Set σ = 4, then generate the original data and stored in matrix X and the response vector in Y .

```
> X
             [,1]       [,2]        [,3]       [,4]       [,5]       [,6]       [,7]        [,8]        [,9]
[1,] 0.631014260 -2.8960374 -3.320506907  1.47641201  3.3433784  4.82775967 -3.0212345 -1.005222269  1.22137718
[2,] 2.129369101 -4.5236801 -4.890305926  1.49692095 -1.2654661 -2.95748005 -3.0163977  0.539677264  4.89372053
[3,] 3.287170948 -4.6766300  3.371366311 -2.19799228 -3.2524430  2.07547786 -2.8439323 -4.839373338  0.16280835
[4,] 1.464932489 -1.9005566  1.393496147  1.68683596 -3.9187363 -3.62870461  0.8699434  4.977291112  3.69914057
[5,] 0.590175737  3.7253480 -0.999094963  4.71276332  2.1728543 -1.60624933  4.3403206  2.983129430 -4.97266536
```

```
> Y
 [1] -54.3014261   5.4720114   3.6983001  41.6853178  13.6897617 -28.2227289  -2.4448810 -44.5834158   7.4265125   5.7941406  -8.7205107  62.2797337
[13]  -5.7273568 -39.1664445 -32.7836809  15.8611099 -13.9127745 -43.2198564 -10.0765997  -2.8355911  69.9472804 -28.5711600 -14.6878878  13.4604027
[25] -37.3574083  -4.4784318  52.7561446  23.4984334  -9.2197777  18.2862849   5.9333964 -14.3365131 -22.8715137  39.5749131  -1.1121141  25.2147882
[37] -61.1747967 -37.6443303   2.4694846 -16.5118219  20.9952744   4.2209804   8.9185209 -46.0510307  45.6448882 -35.9745500  10.7567201 -19.6699527
[49] -11.3226082 -20.8342252  54.6998078  37.2626256 -24.5054246 -26.7468187  -0.8748129 -22.3651593  27.6196649 -14.7816996  -7.6284223  43.4189094
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X5 + X6 + X7, data = df8)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1210 -3.0066  0.3121  2.6430 10.0016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54975    0.28347   1.939   0.0539 .
X1           4.94797    0.09282  53.309   <2e-16 ***
X2           4.88574    0.09857  49.566   <2e-16 ***
X3           1.81446    0.09932  18.269   <2e-16 ***
X5          -4.85635    0.10027 -48.431   <2e-16 ***
X6          -4.97501    0.09867 -50.418   <2e-16 ***
X7          -1.97029    0.09717 -20.277   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.976 on 193 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9817
F-statistic:  1785 on 6 and 193 DF,  p-value: < 2.2e-16
```

i. The superposed plots of the PRESS statistic as a function of the rate of missingness.

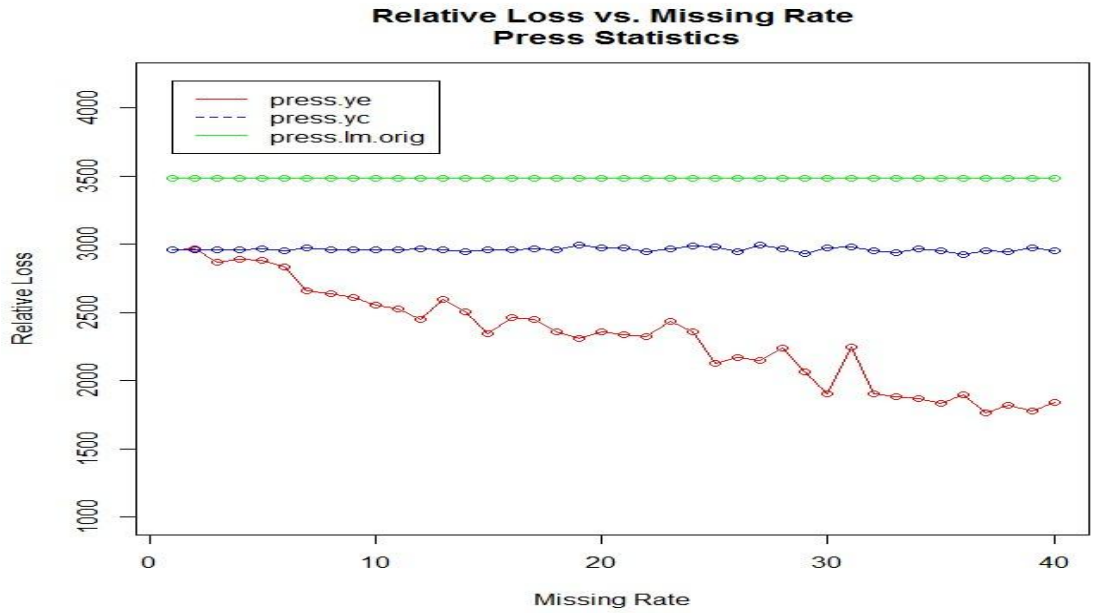

Figure 3.1: Press Statistics setting =

The Press Statistics, when removing NA's, the relative loss appears to decrease significantly as the missingness rate increase, the reason for this is because as the portion of missingness increases the sample size becomes smaller, therefore resulting in a lower press statistics. However, when the missing entries are being imputing using the central tendency the press statistics appears constant and not too extreme from the original Press statistics.

ii. Set σ = 16, then generate the original data and stored in matrix X and the response vector in Y .

```
X
          [,1]       [,2]        [,3]        [,4]        [,5]       [,6]        [,7]        [,8]        [,9]
[1,]  1.15328162 -4.30028356  1.851804582 -1.13000621  3.512500632  2.79539779  0.22785462 -4.459582597  0.02545348
[2,] -1.57052849  4.83541453 -2.259970850  1.54285245 -0.193816777 -3.30312785 -1.90238862 -4.432794005  1.93121085
[3,]  1.19449937 -1.38195675  1.230391711  2.88423790 -3.062228560 -0.94357895  0.29208196  4.019752699  0.76133984
[4,] -3.21040109 -4.72404963 -1.106011283 -0.34382023 -4.411499312  2.72547296  4.29158425  1.363619165 -1.43955120
[5,]  1.36788876 -1.77015949 -2.562507561 -3.71521593 -0.345926494  4.57920997 -3.77890611  0.018305478 -3.59196766

Y
[1] -49.1350906  24.8359435   5.9744059 -39.2253841 -20.5097847  -1.1650569   2.8830139 -32.2492429 -30.9046700   9.6566491   4.8983638 -34.8506072
13] -64.9304418 -27.8639527  37.5805453  54.0697034 -31.0501757 -34.8742116  22.8140981   6.3950739 -46.5643243  43.0949332 -20.8489254 -46.1144026
25]  17.9055359 -20.5455107  37.6033052  -0.8979288 -36.8459225 -12.3013035  34.2007195 -43.6944835 -39.3514069  32.7637878  24.2246428   4.8377526
37]   0.6896268  -1.7525350  13.0566723  10.9921320  27.1843408 -39.4880682  35.3857975   0.7117161 -26.8858738  27.8551232 -57.2221981 -42.9555559
49]   4.6905261  56.4343920  48.1095689  68.2587103 -48.1693348  -6.5727085 -37.7478177  30.6141735  17.2843941   7.3247766  32.8532087 -37.4625857
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X5 + X6 + X7, data = df8)

Residuals:
     Min       1Q   Median       3Q      Max
-43.138   -9.661    0.455    9.360   40.821

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.9959     1.1427  -0.872    0.385
X1             4.8407     0.3836  12.617   < 2e-16 ***
X2             5.3472     0.3982  13.429   < 2e-16 ***
X3             2.2512     0.4009   5.616 6.75e-08 ***
X5            -4.3527     0.3966 -10.974   < 2e-16 ***
X6            -4.3536     0.3904 -11.151   < 2e-16 ***
X7            -1.9043     0.3815  -4.991 1.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.95 on 193 degrees of freedom
Multiple R-squared:  0.7829,    Adjusted R-squared:  0.7762
F-statistic:   116 on 6 and 193 DF,  p-value: < 2.2e-16
```

The superposed plots of the PRESS statistic as a function of the rate of missingness.

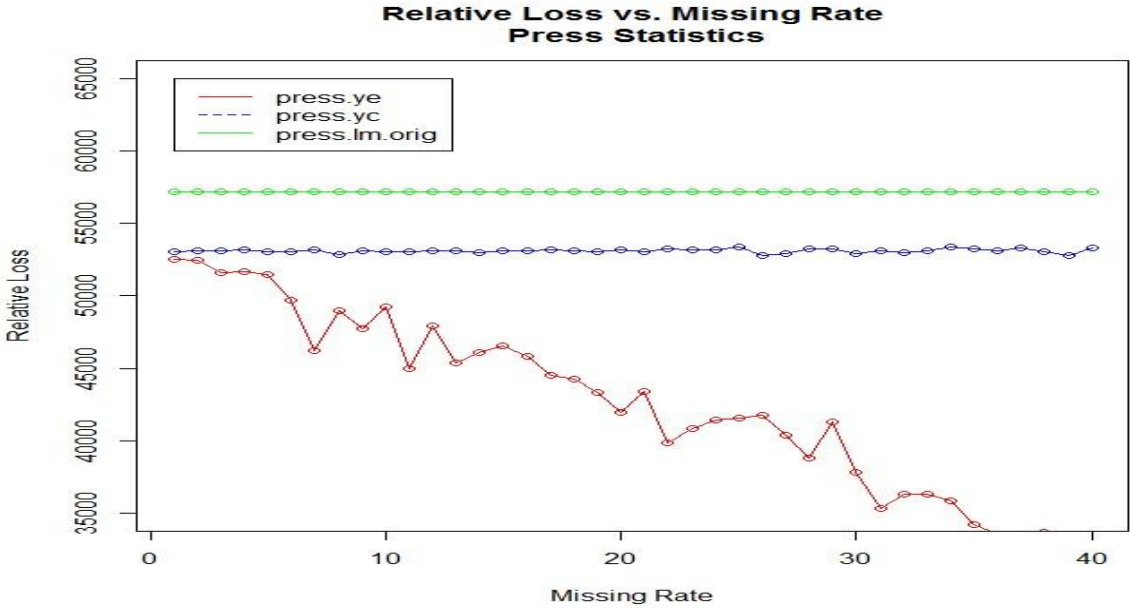

Figure 3.1: Press Statistics setting =

As increases, the scale of press statistics increases. When = ,the press statistics for the original, missingness row deleted and imputation were below 4000, however, as increased ,the press statistics increased tremendously to approximately 55000.

II. CONCLUSION

Although deleting rows with missing entries is a frequent practice by researchers when dealing with missing values in data set, this often results in a substantial decrease in the sample size available for the analysis, which leads to unbiased parameter estimates. From this study it can be concluded that when removing rows with missing entries, there is a significant difference in relative loss in comparison with the technique of central imputation (like mean, median or mode).With the used of generating relative loss using the difference in determinants the method of removing rows with missing entries proven to produce a relative loss up to 100%. The PRESS statistic can be calculated for a number of candidate model structures for the same dataset, with the lowest values of PRESS indicating the best structures. Models that are over-parameterized (over-fitted) would tend to give small residuals for observations included in the model-fitting but large residuals for observations that are excluded. When calculating the PRESS statistics after removing rows with NA's, the relative loss appears to have decline significantly, one may say that this is a great model. However, this can be

justified that this is not a good move and in fact many statisticians/analysis should desist from this method, the reason why the PRESS statistics was shown to be small is because as the rate of missingness increase, the sample size becomes smaller therefore resulting in a lower PRESS statistics.

One point to note is that when variables are highly correlated the used of removing rows with missing entries does not have a negative effect, or one may say that the used of central imputation is not of great power.

Reference:

[1]. Spineli, L.M., Kalyvas, C. Comparison of exclusion, imputation and modelling of missing binary outcome data in frequentist network meta-analysis. BMC Med Res Methodol 20, December 10, 2021 <https://doi.org/10.1186/s12874-020-00929-9>

[2]. Xu, X., Xia, L., Zhang, Q. et al. The ability of different imputation methods for missing values in mental measurement questionnaires. December 13, 2021< https://doi.org/10.1186/s12874-020-00932-0>3.

[3]. Alladoumbaye Ngueilbaye, Hongzhi Wang, Daouda Ahmat Mahamat, Sahalu B. Junaidu,

[4]. Modulo 9 model-based learning for missing data imputation, December 15, 2021 https://doi.org/10.1016/j.asoc.2021.107167

[5]. Weisstein, Eric W. Frobenius Norm. December27,2021 <,https://mathw,orld.wolfram.com/FrobeniusNorm.html >

[6]. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA Correlation Coefficients: Appropriate Use and Interpretation, January 10, 2022

[7]. https://journals.lww.com/anesthesiaanalgesia/fulltext/2018/05000/correlation_coefficiensappropriate_use_and.50.aspx

[8]. Hyun Kang, Korean J Anesthesiol. The prevention and handling of the missing data, January 12, 2022.< https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/>

[9]. Paul Madley-Dowd, Rachael Hughes, Kate Tilling, Jon Heron, The proportion of missing data should not be used to guide decisions on multiple imputation, January 20,2022

[10]. https://doi.org/10.1016/j.jclinepi.2019.02.016.

[11]. Marina Soley-Bori , Dealing with missing data: Key assumptions and methods for applied analysis, February 01,2022, < https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf>

[12]. Ping Xu , The analysis of missing data in public use survey databases : a survey of statistical methods, December 10, 2021 https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=2602&context=etd

[13]. Hyun Kang , The prevention and handling of the missing data , December 01, 2021, < https://www.researchgate.net/publication/237061322_The_prevention_and_handling_of_the_missing_data>

Bibi Sherriza Ali (MSc , BSc), et. al. "Handling Missing Values in Dataset." *IOSR Journal of Mathematics (IOSR-JM)*, 18(2), (2022): pp. 52-62.