Rumainum-HW4

Lince Rumainum

September 29, 2019

Problem 1

Problem 1 (i)

Below is the code in R to create visualization for the three attributes (Na (Sodium), Al (Aluminum), and K (Potassium)) in Glass data that could benefit from a skew transformation using **symbox** function:

```
symbox(Glass$Na, data=Glass, powers=c(3,2,1,0,-0.5,-1,-2))
symbox(Glass$Al, data=Glass, powers=c(3,2,1,0,-0.5,-1,-2))
symbox(Glass$K,  data=Glass, powers=c(3,2,1,0,-0.5,-1,-2))
```

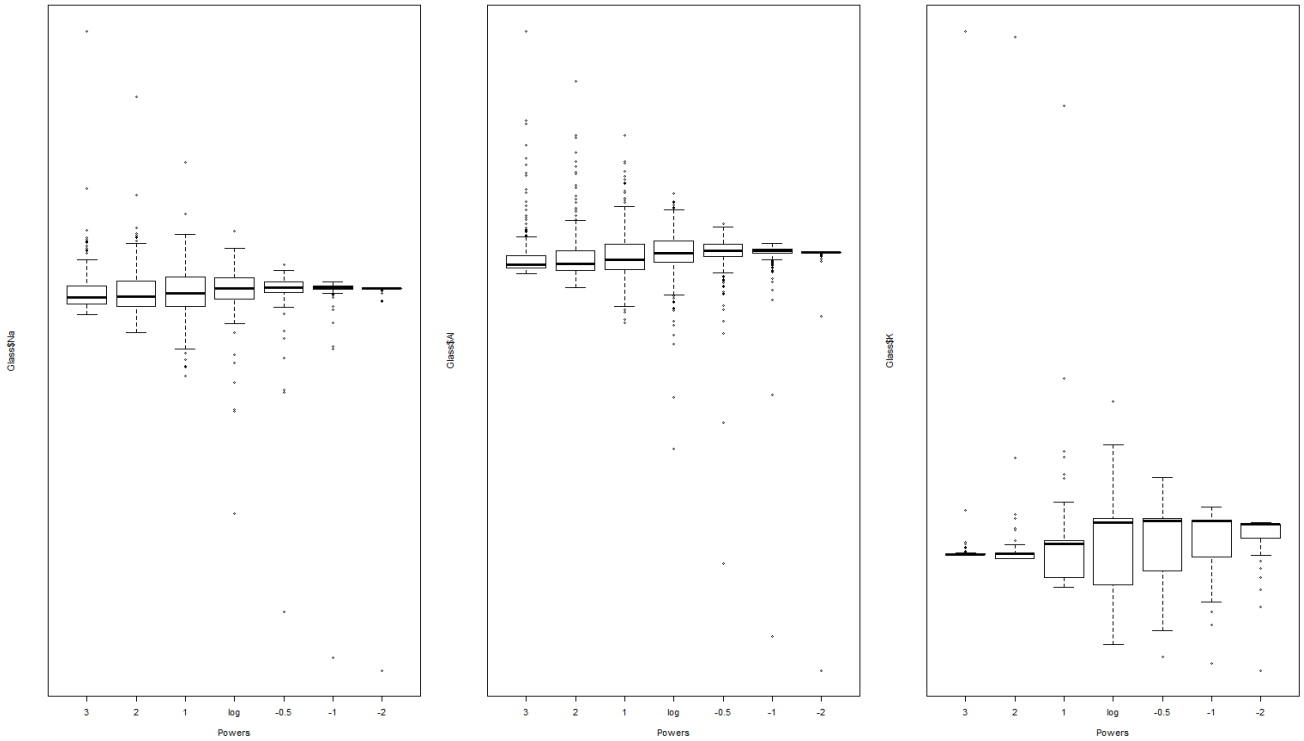

Figure 1 *Visualization of a Skew Tranformation for Na, Al, and K of Glass Dataset*

Problem 1 (ii)

Below are the codes in R to create visualization for the three attributes (Na (Sodium), Al (Aluminum), and K (Potassium)) in Glass data before the transformation, its histogram after it has been transformed, and an updated plot using the **symbox** function with its optimal lambda result from the **boxcox** function:

For Na (Sodium):

```
bc_Na <- boxcox(Glass$Na[Glass$Na>0], optimize = TRUE, lambda=c(-3,3))
bc_Na$lambda
## [1] 0.3795958
par(mfrow=c(1,3))
hist(Glass$Na)
hist((Glass$Na**bc_Na$lambda-1)/bc_Na$lambda)
symbox(Glass$Na, data=Glass, powers=c(2,1,0.3795958,0,-1))
```

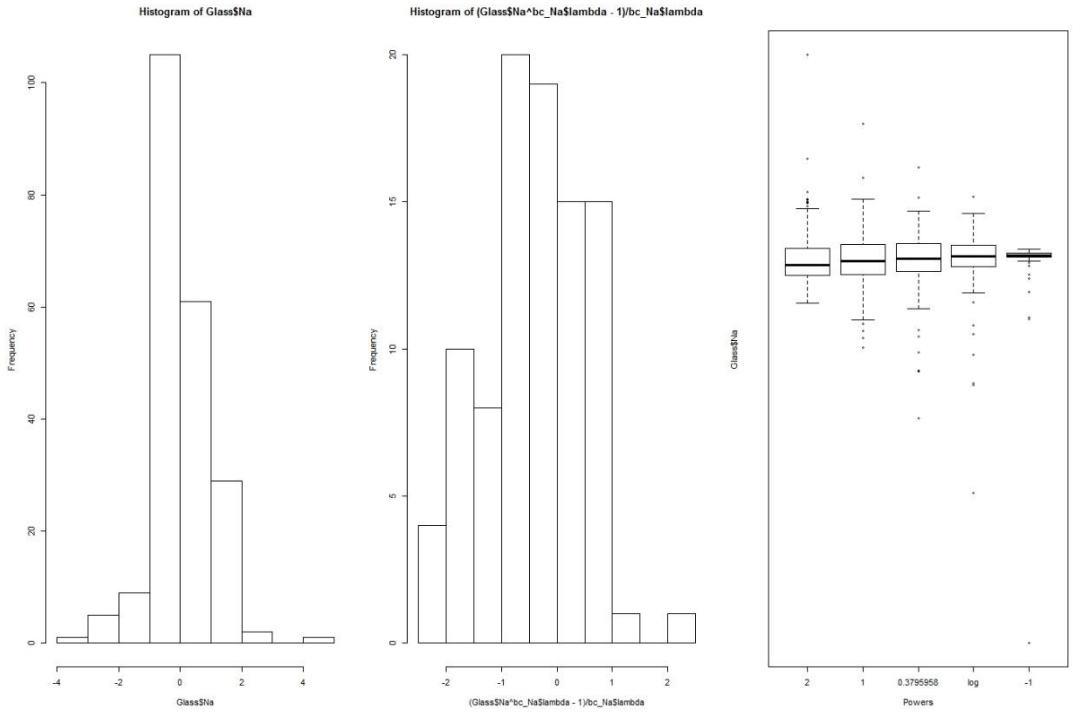

Figure 2 *Histogram of Non-Transform Data, Transform Data, and Updated Symbox Function with Optimal Lambda for Sodium (Na)*

For Al (Aluminum):

```
bc_Al <- boxcox(Glass$Al[Glass$Al>0], optimize = TRUE, lambda=c(-3,3))
bc_Al$lambda
## [1] 0.2357033
par(mfrow=c(1,3))
hist(Glass$Al)
hist((Glass$Al**bc_Al$lambda-1)/bc_Al$lambda)
symbox(Glass$Al, data=Glass, powers=c(2,1,0.2357033,0,-1))
```

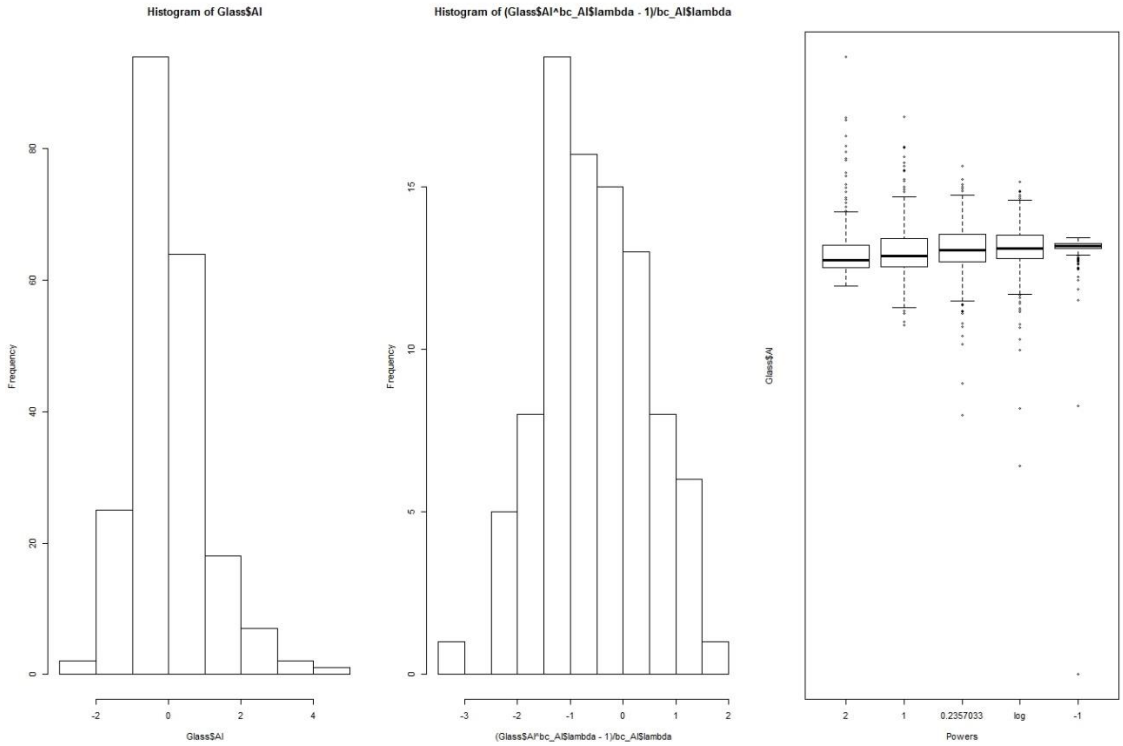

Figure 3 *Histogram of Non-Transform Data, Transform Data, and Updated Symbox Function with Optimal Lambda for Aluminum (Al)*

For K (Potassium):

```
bc_K <- boxcox(Glass$K[Glass$K>0], optimize = TRUE, lambda=c(-3,3))
bc_K$lambda
## [1] -0.03432775
```

```
par(mfrow=c(1,3))
hist(Glass$K)
hist((Glass$K**bc_K$lambda-1)/bc_K$lambda)
symbox(Glass$K, data=Glass, powers=c(2,1,0,-0.03432775,-1))
```

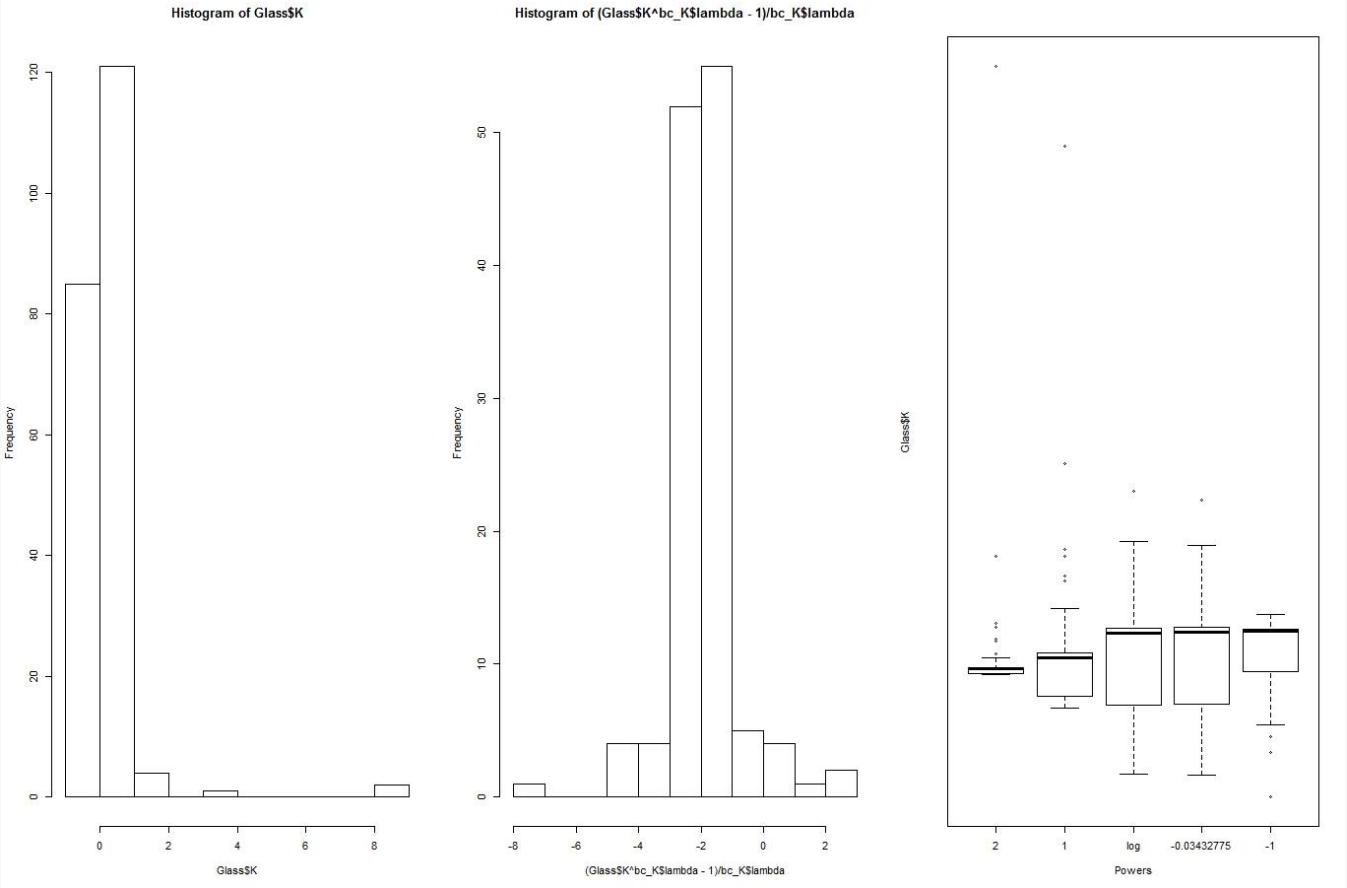

Figure 4 *Histogram of Non-Transform Data, Transform Data, and Updated Symbox Function with Optimal Lambda for Potassium (K)*

Problem 2

Problem 2-a

Exploring the msleep dataset to get an idea of the behavior of the missingness in the data using **aggr** and **md.pattern** function:

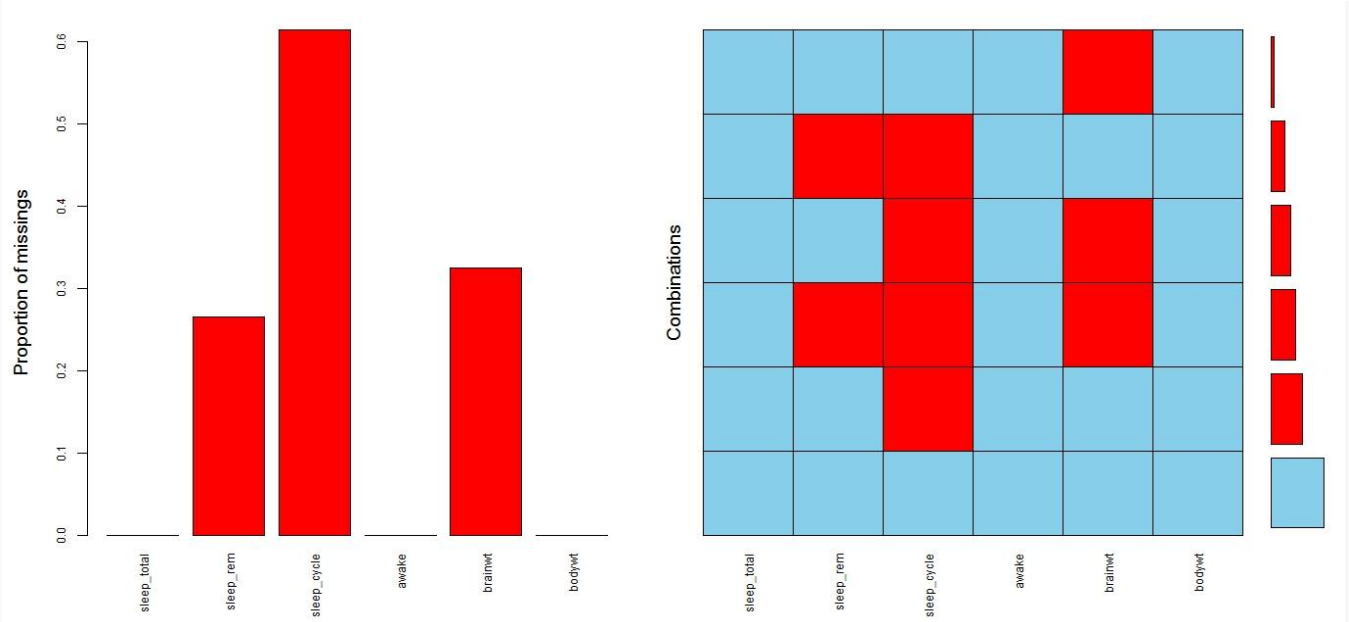

Figure 5 *Missing Data Visualization using aggr Function from VIM package*

```
> md.pattern(msleep_numeric)
   sleep_total awake bodywt sleep_rem brainwt sleep_cycle
30           1     1      1         1       1           1   0
18           1     1      1         1       1           0   1
2            1     1      1         1       0           1   1
11           1     1      1         1       0           0   2
8            1     1      1         0       1           0   2
14           1     1      1         0       0           0   3
             0     0      0        22      27          51 100
```

Figure 6 *Missing Data using md.pattern from mice Package*

The pattern shows that there are 18 instances where sleep_cycle is missing, two instances where brain weight is missing, eleven instances where brain weight and sleep cycle is missing, eight instances where REM sleep time and sleep cycle are missing, and fourteen instances where REM sleep time, brain weight, and sleep cycle data are missing. To summarize, there are 26.5% missing data in sleep_rem, 61.4% missing in sleep_cycle, and 32.5% missing in brain weight while total sleep time, awake time, and body weight have complete data.

Problem 2-b

The article by Savage and West (2007) is using the scaling relationships of simple power laws. They are using the ladder of powers by transforming their data into ln-ln space. Since the variance relations and relative error between the attributes that they analyzed are linear, the predicted regression data were done using ordinary least squares regression method. They analyze the relation between the ratio of the sleep time and awake time with the body mass and also with brain mass and the ratio of REM sleep time and the total sleep time with their body mass.
For this problem, before transforming the data doing the standardization of the data is important so all the data will be in the same scale. In this case, z-score standardization will be used. Since some attributes are heavily skewed, transforming skewness will done by using ladder of powers method using the **boxcox** function in R.

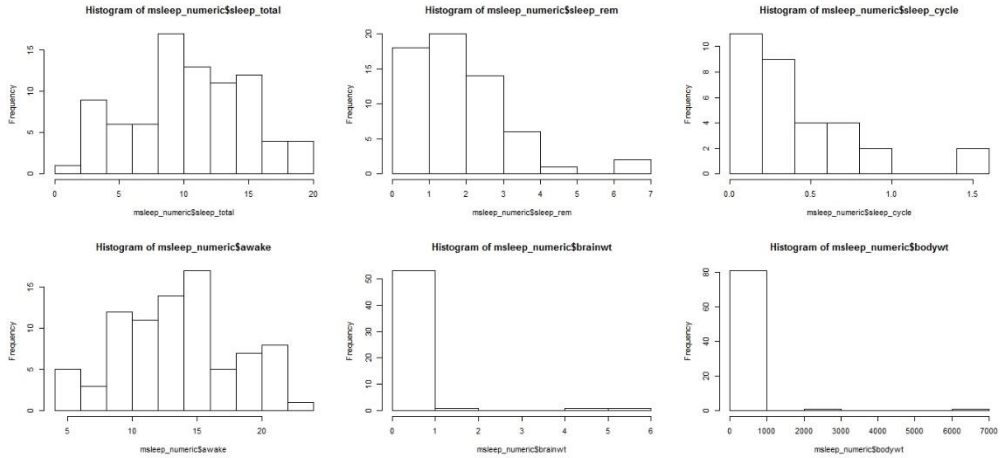

Figure 7 *Histogram of msleep Data Before Transformation*

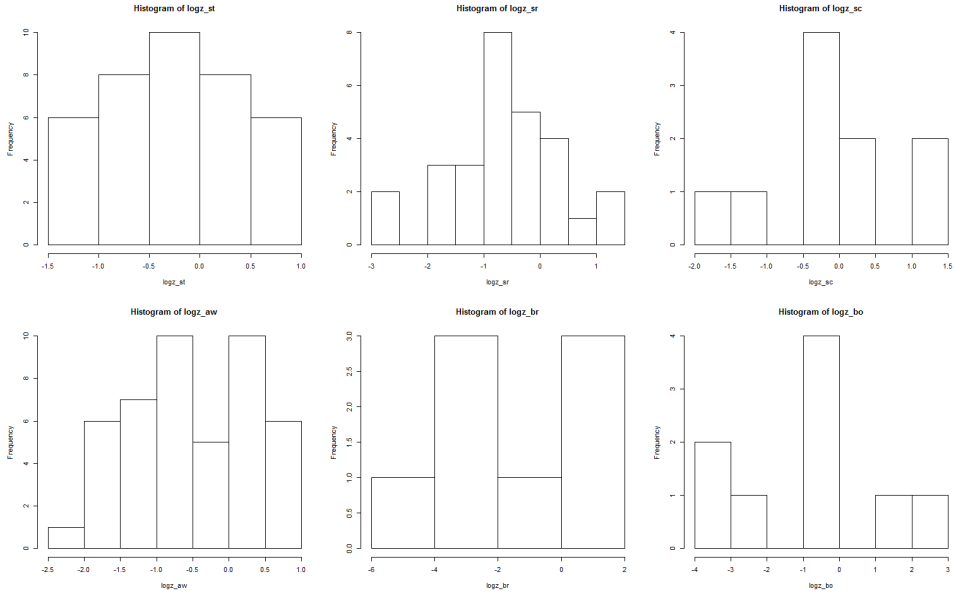

Figure 8 *Histogram of msleep Data after Standarized to Z-Scores*

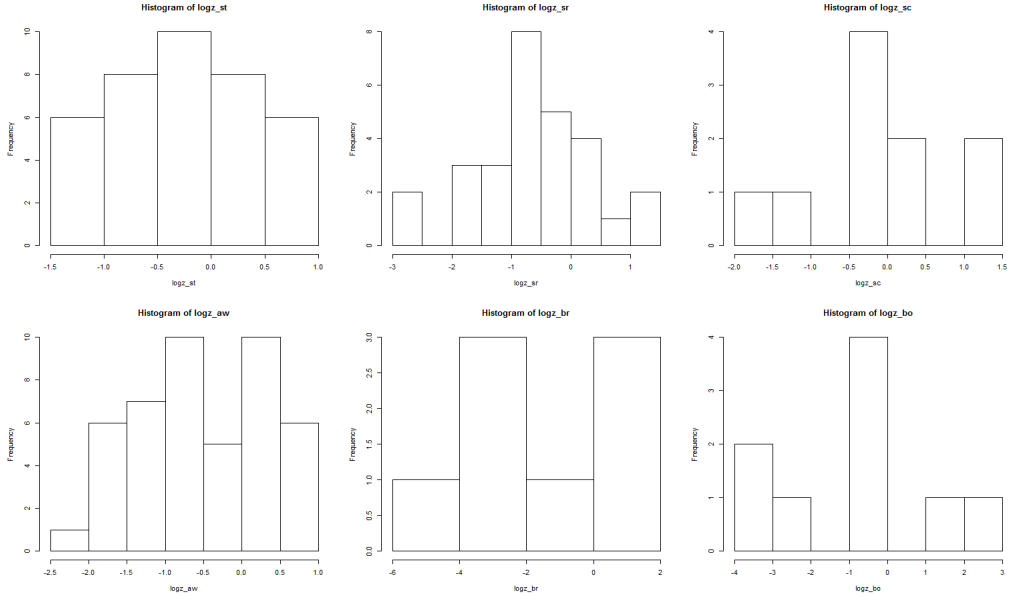

Figure 9 *Histogram of msleep Data after being Skewed Transformed using Each Attribute's Optimal Lamda*

Below shows the relationship between body weight, brain weight, and the total sleep when group by the order type of mammals:

order	n	brainwt	bodywt	sleep_total
<chr>	<int>	<dbl>	<dbl>	<dbl>
1 Proboscidea	2	5.16	4600.	3.6
2 Perissodactyla	3	0.414	305.	3.47
3 Primates	12	0.254	13.9	10.5
4 Artiodactyla	6	0.198	282.	4.52
5 Carnivora	12	0.0986	57.7	10.1
6 Cingulata	2	0.0459	31.8	17.8
7 Monotremata	1	0.025	4.5	8.6
8 Hyracoidea	3	0.0152	3.06	5.67
9 Lagomorpha	1	0.0121	2.5	8.4
10 Diprotodontia	2	0.0114	1.36	12.4
11 Didelphimorphia	2	0.0063	1.03	18.7
12 Rodentia	22	0.00357	0.288	12.5
13 Erinaceomorpha	2	0.00295	0.66	10.2
14 Afrosoricida	1	0.0026	0.9	15.6
15 Scandentia	1	0.0025	0.104	8.9
16 Soricomorpha	5	0.000592	0.0414	11.1
17 Chiroptera	2	0.000275	0.0165	19.8
18 Cetacea	3	NaN	342.00	4.5
19 Pilosa	1	NaN	3.85	14.4

Below shows a linear relationship between brain weight, total sleep time, sleep cycle, and the REM sleep time as the range of mamal's total body weight increases:

bodywt_group	n	brainwt	sleep_total	sleep_cycle	sleep_rem
<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1 [0,1)	35	3.66	12.8	0.205	2.3
2 [1,5)	20	19.2	10.7	0.444	1.81
3 [5,10)	1	179	10.1	0.75	1.2
4 [10,50)	6	116.	7.9	0.461	1.4
5 [50,100)	9	420.	8.46	1.14	2.04
6 [100,1e+03)	10	365.	6.39	0.856	0.533
7 [1e+03,Inf)	2	5158.	3.6	NaN	NaN

Problem 2-c (i)

For part (i), the method used is linear regression ignoring model error, where the model creates eight of multiple imputations and impute missing values for 100 iterations for all the missing numeric fields (REM sleep time, sleep cycle, and brain weight).

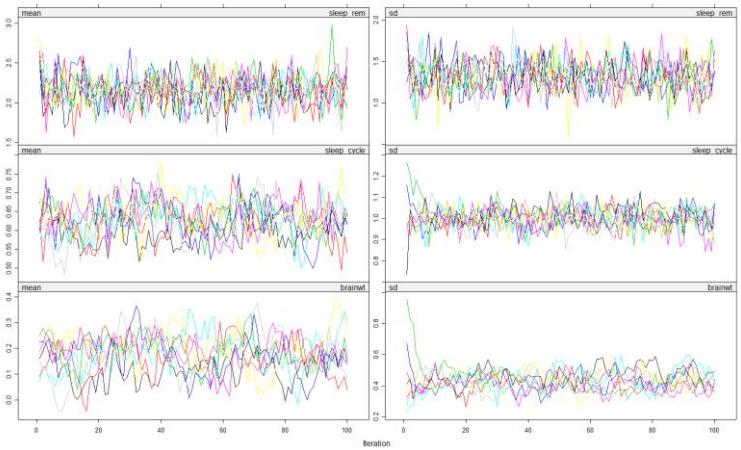

Figure 10 *Shows the Mean and Variance Behavior for Each Iteration with* ***Linear Regression Ignoring Model Error*** *Model for Each Missing Attribute*

Problem 2-c (ii)

Using the **with** and **pool** function of mice package, the imputed data is built and evaluated based on a linear regression model. The **with** function takes all of the eight copied datasets to the fit model, which based on linear regression of all the numeric attributes. In this model, random noise is introduced to help avoid overfitting. Once that process is done, the **pool** function will recombining all of those data to create the best estimate for the missing values. Below are the R codes used to do the **with** and **pool** function:

```
fit_msleep <- with(imp_msleep, lm(bd~st+sr+sc+aw+br+randNoise))
est_msleep <- pool(fit_msleep)
```

Problem 2-c (iii)

The results of regressions coefficients and p-values of the same linear regression model on complete numeric cases *without* missing value:

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.99  -60.15  -22.93   18.45  366.76
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  257.487     97.134   2.651   0.0140 *
## st           -16.865      7.895  -2.136   0.0431 *
## sr            -6.672     28.674  -0.233   0.8180
## sc           -62.730    121.953  -0.514   0.6117
## aw                NA         NA      NA       NA
## br           197.358    155.102   1.272   0.2154
## randNoise     -7.915     27.767  -0.285   0.7781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 120.3 on 24 degrees of freedom
##   (53 observations deleted due to missingness)
## Multiple R-squared:  0.4259, Adjusted R-squared:  0.3063
## F-statistic: 3.561 on 5 and 24 DF,  p-value: 0.01502
```

The results of regressions coefficients and p-values of the same linear regression model on complete numeric cases *with* missing value:

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.99  -60.15  -22.93   18.45  366.76
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  257.487     97.134   2.651   0.0140 *
## st           -16.865      7.895  -2.136   0.0431 *
## sr            -6.672     28.674  -0.233   0.8180
## sc           -62.730    121.953  -0.514   0.6117
## aw                NA         NA      NA       NA
## br           197.358    155.102   1.272   0.2154
## randNoise     -7.915     27.767  -0.285   0.7781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 120.3 on 24 degrees of freedom
##   (53 observations deleted due to missingness)
## Multiple R-squared:  0.4259, Adjusted R-squared:  0.3063
## F-statistic: 3.561 on 5 and 24 DF,  p-value: 0.01502
```

The results of regressions coefficients and p-values of the same linear regression model on the estimates based on the linear regression model:

```
##                estimate  std.error  statistic    df    p.value
## (Intercept) 257.487256  97.133637  2.6508557 22.22 0.01452607
## st          -16.865392   7.895067 -2.1361937 22.22 0.04392270
## sr           -6.672494  28.674015 -0.2327018 22.22 0.81812309
## sc          -62.730126 121.953490 -0.5143775 22.22 0.61206654
## br          197.358468 155.101905  1.2724439 22.22 0.21637024
## randNoise    -7.914750  27.767433 -0.2850372 22.22 0.77825555
```

Problem 2-c (iv)

For part (i), the method used is linear regression, predicted values, where the model creates eight of multiple imputations and impute missing values for 50 iterations for all the missing numeric fields (REM sleep time, sleep cycle, and brain weight).

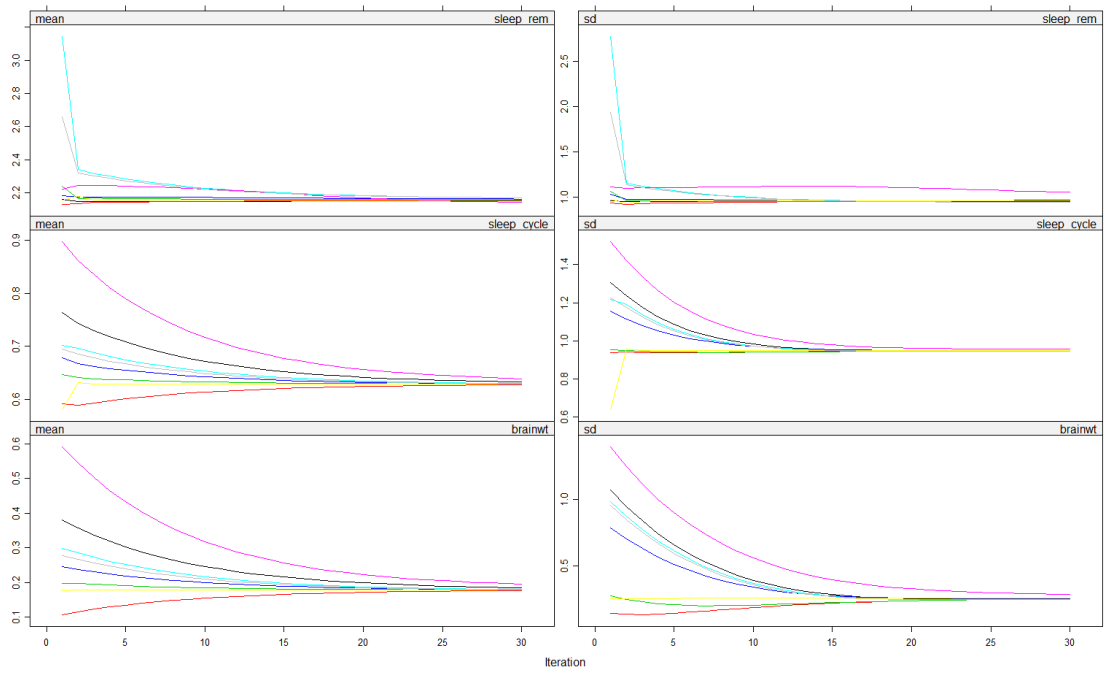

Figure 11 *Shows the Mean and Variance Behavior for Each Iteration with* ***Linear Regression, Predicted Values*** *Model for Each Missing Attribute*

The results of regressions coefficients and p-values of the same linear regression model on the estimates based on the linear regression model:

```
##                estimate  std.error  statistic    df    p.value
## (Intercept) 257.487256  97.133637  2.6508557 22.22 0.01452607
## st          -16.865392   7.895067 -2.1361937 22.22 0.04392270
## sr           -6.672494  28.674015 -0.2327018 22.22 0.81812309
## sc          -62.730126 121.953490 -0.5143775 22.22 0.61206654
## br          197.358468 155.101905  1.2724439 22.22 0.21637024
## randNoise    -7.914750  27.767433 -0.2850372 22.22 0.77825555
```

From the two models, ***Linear Regression Ignoring Model Error*** and ***Linear Regression, Predicted Values***, it shows that ***Linear Regression, Predicted Values*** produce better results because the behavior of the mean and standard deviation with half of the iterations converging while ***Linear Regression Ignoring Model Error*** model does not.