STAT270 Assignment 2

Session 2 2017

Justin Lam 45197083

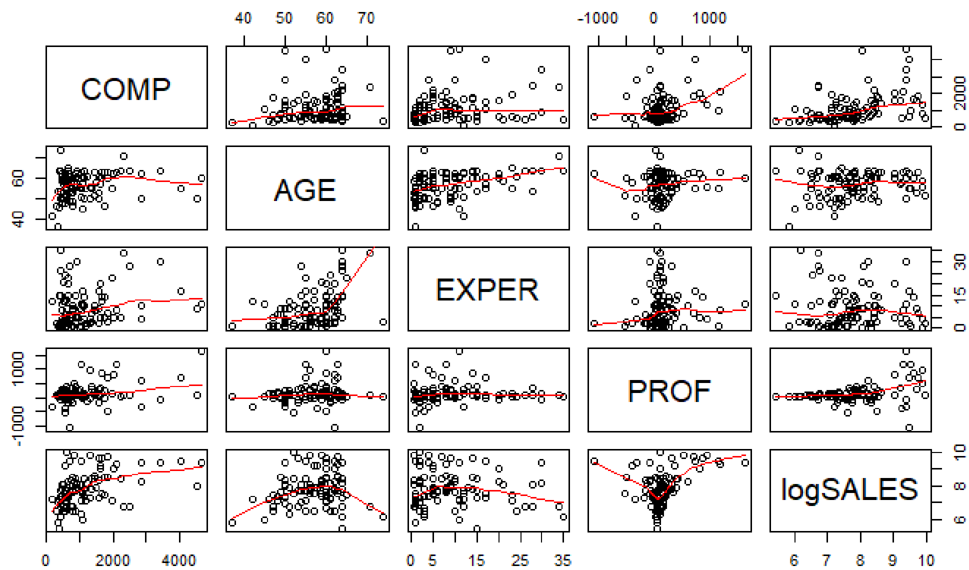
Due: 3rd November 2017 5:00pm

Question 1

a) CEO = read.csv("CeoCompensation.csv", header = TRUE

Plot (CEO)

pairs (CEO, panel = panel.smooth)



By observing the scatter plots above, it demonstrates the relationship between the response and predictors variables. Firstly, there is a positive relationship between COMP (as a response variable) and the all its predictor variables (such as AGE, EXP, logSALES and PROF). In addition, the correlation of between COMP and variables PROF and logSALES seems to be greater than between COMP, AGE and EXPER. Secondly, the correlation between AGE and EXP is strongest between AGE (as a response variable) and other predictor variables; they are COMP, logSALES and PROF. There is almost no relationship between them. In addition, it seems to be a negative correlation between EXPER and logSALES. Thirdly, we can interpret the correlation of EXPER (as a response variables) and other variables being the predictor variables.

Notably, the correlation between EXPER, COMP and PROF are the strongest and the rest of the variables are relatively weak. Moreover, we also need to examine the correlation between PROF (as a response variable) and other variables being a predictor variables. The correlation between PROF, COMP and logSALES are significant than the others even though they are only showing a weak positive relationship. Finally, we also need determine the correlation between logSALES (being a response variable) and other variables being the predictor variable. Generally, the correlation between logSALES and all variables have a weak positive relationship; except logSALES and EXP which appears to have a weak negative variable. Given the fact that the correlation is weak within all variables therefore it may impact the accuracy and creditability of the data set. As a result, transformation of data (such as a log transform or a square root transform) is required to repair such problem.

b) cor (CEO) – a command which allows R to compute the correlation matrix of the dataset.

```
> cor(CEO)
              COMP       AGE       EXPER       PROF    logSALES
COMP     1.0000000 0.1523392  0.22599438 0.37502816  0.43590914
AGE      0.1523392 1.0000000  0.40765174 0.13051350  0.12643284
EXPER    0.2259944 0.4076517  1.00000000 0.07614768 -0.06408743
PROF     0.3750282 0.1305135  0.07614768 1.00000000  0.35022687
logSALES 0.4359091 0.1264328 -0.06408743 0.35022687  1.00000000
```

c) Conducting an F-test for the multiple regression model:

CEO.lm = lm (COMP ~ AGE + EXPER + logSALES + PROF, data = CEO) – defining the linear model CEO (start with all the variables)

summary (CEO.lm) – computes a numerical summary of the linear model.

anova (CEO.lm) – conduct ANOVA to get both the F-statistics and the p-value.

```
> anova(CEO.lm)
Analysis of Variance Table

Response: COMP
          Df   Sum Sq  Mean Sq F value    Pr(>F)    
AGE        1  1670609  1670609  3.1498   0.07914 .  
EXPER      1  2318998  2318998  4.3722   0.03920 *  
logSALES   1 14356802 14356802 27.0682 1.129e-06 ***
PROF       1  3252805  3252805  6.1328   0.01504 *  
Residuals 95 50387386   530394                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

The multiple regression model and its parameters:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + \varepsilon$

β_0: Intercept term

$\beta_1 \ldots \beta_k$: partial regression coefficients:

ε: random standard error

Hypothesis Testing:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: at least one of the β_i is not equal to 0

The ANOVA table is shown above (underneath the ANOVA command)

F-statistic can be conducted by the command: summary (CEO.lm) in R, in the output provided, it shows the F-statistic is 10.18 with 4 and 95 degrees of freedom.

The null distribution of the test = $F_{4, 95}$

P-value = 6.671e-07 (obtained from the summary command @R)

Statistical Conclusion: Given the fact the p-value is < 0.05, therefore we will reject H_0.

Contextual conclusion: Since we reject the H0, therefore it suggests that the age (the CEO"s age), has no effect in COMP (sum of salary, bonus and other compensation).

d) In the question, it will demonstrate the process of backward model selection procedure to find the best multiple regression model which explains the data by using COMP as a response variable and start will all the predictors provided.

Firstly, we will repeat the process of conducing ANOVA in order to determine which predictor is insignificant.

```
> anova(CEO.lm)
Analysis of Variance Table

Response: COMP
          Df   Sum Sq  Mean Sq F value    Pr(>F)
AGE        1  1670609  1670609  3.1498   0.07914 .
EXPER      1  2318998  2318998  4.3722   0.03920 *
logSALES   1 14356802 14356802 27.0682 1.129e-06 ***
PROF       1  3252805  3252805  6.1328   0.01504 *
Residuals 95 50387386   530394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

From the output above, it shows that "AGE" predictor is insignificant due to the fact that the p-value of such predictor is 0.07914 and it is larger than the significance level and hence we need to remove the "AGE" predictor from this model.

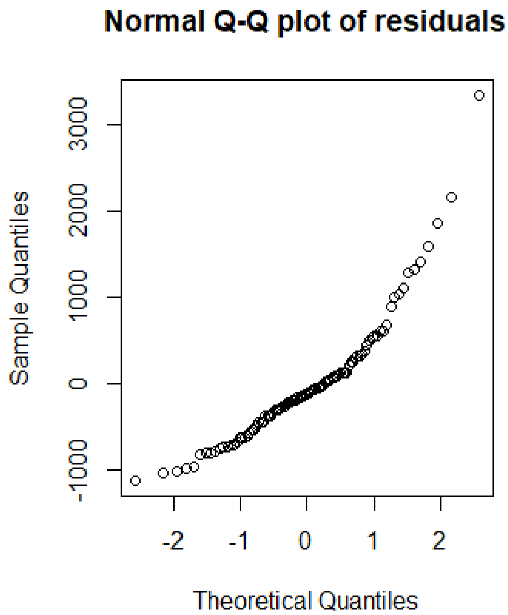
```
Call:
lm(formula = COMP ~ EXPER + logSALES + PROF, data = CEO)

Residuals:
    Min      1Q  Median      3Q     Max
-1084.0  -449.5  -115.5   237.3  3357.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1641.8771   616.3384  -2.664 0.009059 **
EXPER          23.9670     8.8736   2.701 0.008175 **
logSALES      316.6802    78.1693   4.051 0.000103 ***
PROF            0.5660     0.2283   2.479 0.014923 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 724.7 on 96 degrees of freedom
Multiple R-squared:  0.2996,	Adjusted R-squared:  0.2777
F-statistic: 13.69 on 3 and 96 DF,  p-value: 1.665e-07
```

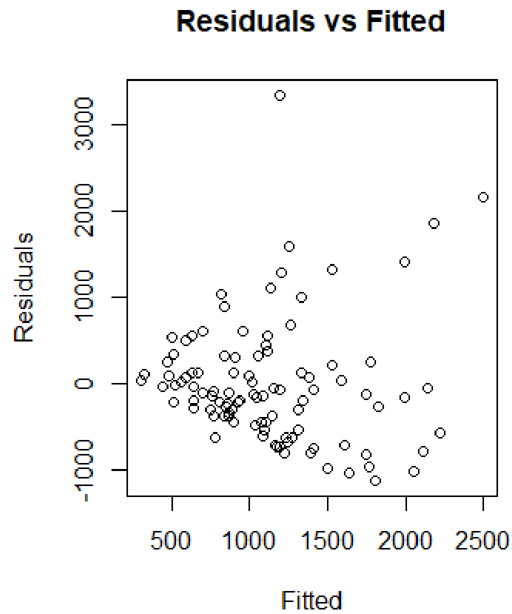
The above diagram shows the overall ANOVA table, after the "AGE" predictor has been received. By observing the p-values of other predictors, it seems that they are significant as they are all less than the significance level (0.05). Overall, backward model selection procedure has been used (by removing the predictor "AGE") in this question and the best multiple regression model has been found.

e) In order to validate the model to explain why it is not appropriate to use multiple regression model (to explain the COMP response), we can use the assumptions of the multiple regression model to do so. Recalling from the lectures, the two assumptions used in validate the multiple regression model are the residuals vs fitted plot (to determine whether there's an obvious relationship or not) and the normal Q-Q (Quantile-Quantile) plot where we use to determine the normality

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1487.3419   881.0902  -1.688 0.094678 .
AGE            -3.2502    13.1797  -0.247 0.805747
EXPER          24.9679     9.7977   2.548 0.012427 *
logSALES      319.3854    79.3167   4.027 0.000114 ***
PROF            0.5691     0.2298   2.476 0.015038 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

assumption of the model. The plots are shown below:





From the above diagrams, it shows the assumption component of the hypothesis test. Even though there is no obvious pattern in the "residuals vs fitted" diagram, nevertheless, in "Normal Q-Q plot of residuals", the normality assumption has been violated. This is due to the outliers shown in the plot and hence it is not appropriate to use the multiple regression model to explain the COMP response.

f) The following question will refit the multiple regression by using log(COMP) as the new predictor variable. In addition, backward selection procedure will apply in this question.

```
Call:
lm(formula = log(COMP) ~ AGE + EXPER + logSALES + PROF, data = CEO)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.35067 -0.38023 -0.03568  0.33445  1.62218 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.0047956  0.6435444   6.223 1.31e-08 ***
AGE         0.0088861  0.0096264   0.923   0.3583    
EXPER       0.0134722  0.0071562   1.883   0.0628 .  
logSALES    0.2730743  0.0579326   4.714 8.32e-06 ***
PROF        0.0003336  0.0001678   1.987   0.0498 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5319 on 95 degrees of freedom
Multiple R-squared:  0.3292,    Adjusted R-squared:  0.301 
F-statistic: 11.66 on 4 and 95 DF,  p-value: 9.622e-08
```

As shown in the above image, log (COMP) has become the new predictor variable and all the dependent variables are all shown in the formula. It also shows that the 'AGE" and "EXPER' are both insignificant, as their p-values are both > 0.05 and thus backward model selection is required to remove them.

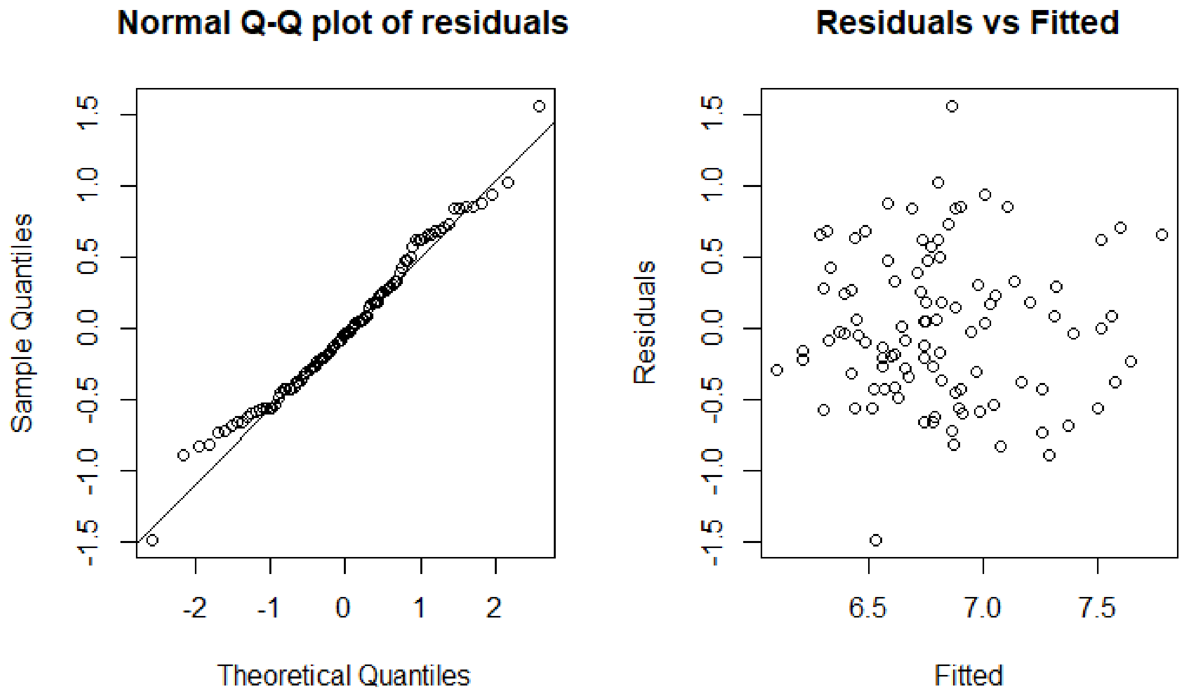
Firstly, 'AGE" variable will be dropped.

```
Call:
lm(formula = log(COMP) ~ EXPER + logSALES + PROF, data = CEO)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48328 -0.38838 -0.03446  0.32870  1.55913 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.4273026  0.4520406   9.794 4.12e-16 ***
EXPER       0.0162089  0.0065082   2.491   0.0145 *  
logSALES    0.2804706  0.0573316   4.892 4.01e-06 ***
PROF        0.0003419  0.0001675   2.042   0.0439 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5315 on 96 degrees of freedom
Multiple R-squared:  0.3232,	Adjusted R-squared:  0.3021 
F-statistic: 15.28 on 3 and 96 DF,  p-value: 3.325e-08
```

The above image shows the effect of removing "AGE' variable and the linear Model. By observing the p-value (to determine which variables are significant/insignificant), it seems all the variables are significant as their p-values are all < 0.05 and hence it is the final model.

g) Lastly, this question will validate the final model with log(COMP) response and it will also explain why the regression model with log(COMP) response variable is superior to the model with the COMP response variable.

Firstly, the above question has already shown the final model with log (COMP) response which is 'lm (formula = log (COMP) ~ EXPER + logSALES + PROF, dat a = CEO)'. The F-statistic and the p-value are also shown in the above output. In addition, we can also compare the residuals vs fitted plot and the normal Q-Q plot (the assumptions of the multiple regression) to explain why the log(COMP) is superior than the normal COMP response.

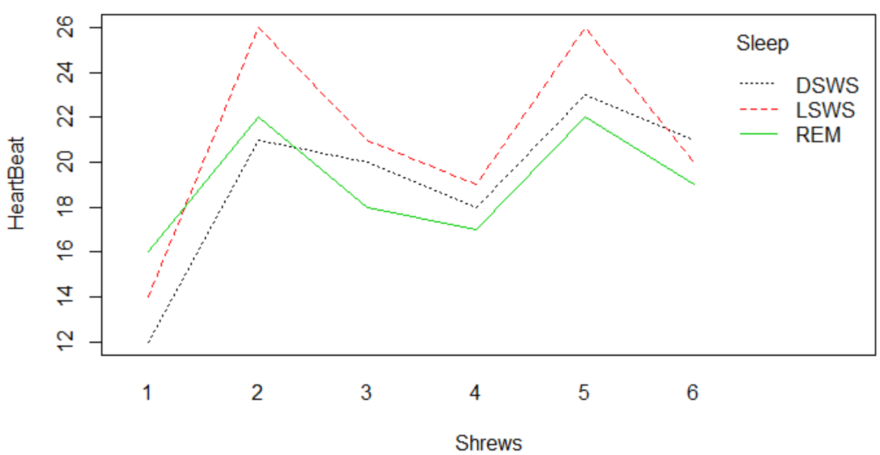


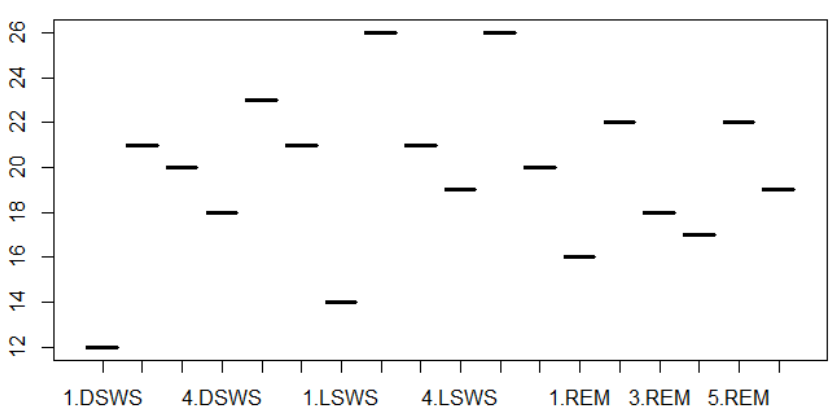
The above image shows the normal Q-Q plot and the residuals vs fitted plot for the log(COMP) response. By comparing these models the one listed at page 4 (Which is the COMP response), the normality assumption within the log (COMP) model is better as this can be shown in the attached Q-Q plot. Furthermore, there is no obvious patterns with the residuals vs fitted and thus the assumptions are valid for the model. Overall, by comparing the multiple regression assumptions of both log (COMP) and COMP response, it clarifies the reason why log (COMP) response is superior to the other response.

Question 2

a) The design (TreeStrews.dat.txt) is balanced because it has the same number of group sizes (replicates).

b) The plots below shows an interaction ploy and a box plot and they are the preliminary graphs which are used to investigate the features of the data. Firstly, there is a very strong interaction between HeartBeats being the variable, Shrews on the axis and Sleep being the other factor that controls the line type in the model. Since there is a strong interaction, therefore the change in response to Shrews and will not be the same as the level of Sleep. Moreover, we can also use boxplots to visualise the effects of the variability of the data. From the output below, it shows that there is a lack of arability in the data set due to the variation sample sizes. Overall, the above response has explained the two preliminary graphs that used to investigate the features of the data.





c) We cannot fit a two-way ANOVA model in this situation due the fact the sample size of residuals is 0. This can also be shown in the R output below which it highlights an error message to user.

```
Analysis of Variance Table

Response: HeartRates
                     Df  Sum Sq Mean Sq F value Pr(>F)
factor(Shrews)        5 186.278  37.256
Sleep                 2  14.778   7.389
factor(Shrews):Sleep 10  24.556   2.456
Residuals             0   0.000
Warning message:
In anova.lm(TreeShrewslm) :
  ANOVA F-tests on an essentially perfect fit are unreliable
> 
```

d) The mathematical model for two-way ANOVA = $Y_{ij}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\epsilon_{ijk}+\epsilon$

The parameters are listed as follows:

- **Response:** γ_{ij} = kth replicate of the treatment at i^{th} level in factor A and ^{jth} level in factor B.
- μ: overall population mean.
- α_i: base effect of i^{th} level of Factor A; i = 1,2….,a.
- βj: base effect of jth level of Factor B; j = 1,2....,b.
- $\gamma_{ij:}$ effect of combined effect of the i^{th}, j^{th} combination the two factors.
- ϵ_{ijk}: unexplained variation for each replicated observation $\epsilon \sim \mathcal{N}$ $(0, \sigma^2)$

Null Hypothesis (H_0):

1. The population mean of the factor “Shrews” are equal.
2. The population mean of the factor “Sleep” are equal.
3. There is no interaction between the “Shrews” and “Sleep” factor.

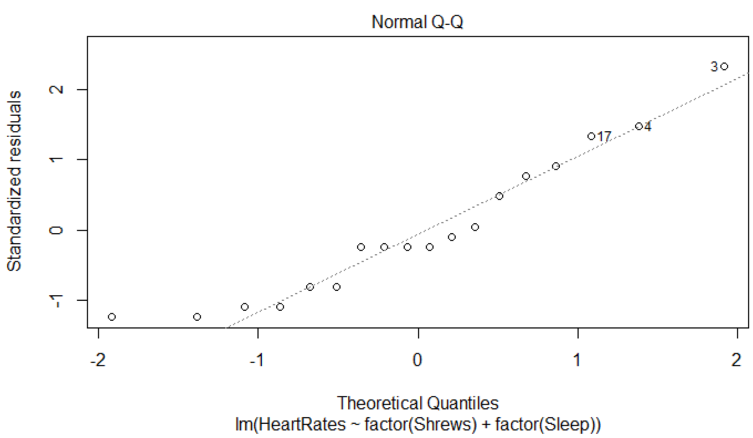
Alternative hypothesis (H_1):

At least one of the statements in the null hypothesis is invalid.

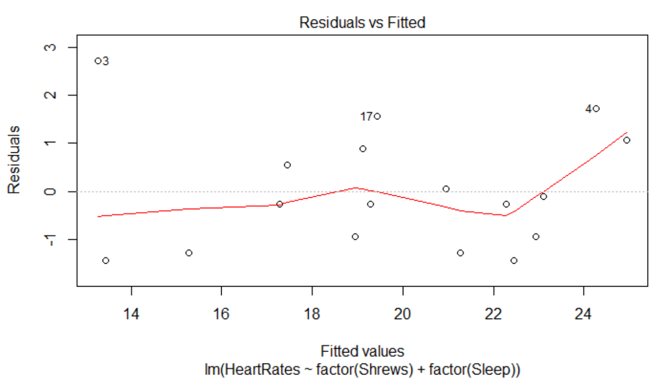
ANOVA Table (removing the interaction terms as it disallows us to conduct an ANOVA test)

```
Response: HeartRates
               Df  Sum Sq Mean Sq F value    Pr(>F)    
factor(Shrews)  5 186.278  37.256  15.172 0.0002157 ***
factor(Sleep)   2  14.778   7.389   3.009 0.0948298 .  
Residuals      10  24.556   2.456                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can check assumptions by using both Q-Q plot and Residuals vs Fitted plot.



Moreover, the output from the Q-Q plot is somewhat linear and therefore the assumption has been met for two-way ANOVA.



Since there is no obvious (clear) relationship within the Residuals vs Fitted plot and thus the assumption of two-way ANOVA is valid.

e) Conclusion. Since the p-value of factor (Shrews) is smaller than 0.05, therefore it is significant indicator to predict the heartrates of tree shrew. On the other hand, the p-value of factor (Sleep) is greater than 0.05 therefore it is insignificant to the heartbeats of tree shrew.