HW1 Peer Assessment

Part A. ANOVA

Additional Material: ANOVA tutorial

https://datascienceplus.com/one-way-anova-in-r/ (https://datascienceplus.com/one-way-anova-in-r/)

Jet lag is a common problem for people traveling across multiple time zones, but people can gradually adjust to the new time zone since the exposure of the shifted light schedule to their eyes can resets the internal circadian rhythm in a process called “phase shift”. Campbell and Murphy (1998) in a highly controversial study reported that the human circadian clock can also be reset by only exposing the back of the knee to light, with some hailing this as a major discovery and others challenging aspects of the experimental design. The table below is taken from a later experiment by Wright and Czeisler (2002) that re-examined the phenomenon. The new experiment measured circadian rhythm through the daily cycle of melatonin production in 22 subjects randomly assigned to one of three light treatments. Subjects were woken from sleep and for three hours were exposed to bright lights applied to the eyes only, to the knees only or to neither (control group). The effects of treatment to the circadian rhythm were measured two days later by the magnitude of phase shift (measured in hours) in each subject’s daily cycle of melatonin production. A negative measurement indicates a delay in melatonin production, a predicted effect of light treatment, while a positive number indicates an advance.

Raw data of phase shift, in hours, for the circadian rhythm experiment

Treatment	**Phase Shift (hr)**
Control	0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27
Knees	0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61
Eyes	-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83

Question A1 - 3 pts

I am going to copy the above row data in to the text file and load that text file here

```
phase_data <- read.table("hw1.txt",sep = "|",header = T)
s <- strsplit(phase_data$PhaseShift, split = ",")
s<- data.frame(Treatment = rep(phase_data$Treatment, sapply(s, length)), PhaseShift = unl
        ist(s))
s$PhaseShift <- as.numeric(s$PhaseShift)
head(s)
```

```
##   Treatment PhaseShift
## 1   Control       0.53
## 2   Control       0.36
## 3   Control       0.20
## 4   Control      -0.37
## 5   Control      -0.60
## 6   Control      -0.64
```

Let's build the ANOVA table.....

```
phase_model <- aov(PhaseShift~Treatment,data = s)
summary(phase_model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment    2  7.224   3.612   7.289 0.00447 **
## Residuals   19  9.415   0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above table we can fill in the below table.

Consider the following incomplete R output:

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	2	7.224	3.6122	7.289	0.004
Error	19	9.415	0.496		
TOTAL	21	16.639			

Fill in the missing values in the analysis of the variance table.

Question A2 - 3 pts

Use μ_1, μ_2, and μ_3 as notation for the three mean parameters and define these parameters clearly based on the context of the topic above. Find the estimates of these parameters.
μ_1:Control: -0.3087
μ_2:Eyes: -1.1551
μ_3:Knees: -0.3357

I am going to calculate mean by using below code

```
model.tables(phase_model,type = "means")
```

```
## Tables of means
## Grand mean
##
## -0.7127273
##
##  Treatment
##     Control   Eyes   Knees
##     -0.3087 -1.551 -0.3357
## rep  8.0000  7.000  7.0000
```

Question A3 - 5 pts

Use the ANOVA table in Question A1 to answer the following questions:

a. **1 pts** Write the null hypothesis of the ANOVA F-test, H_0
   # H_0: The Variances are likely to be equal. That means Mean Square(between)= Mean Square(within) or
   Mean Square(between)/Mean Squar(within) = 1

b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A
   # H_A : The Variances are likely to be different. That means Mean Square(between) != Mean Square(within) or
   Mean Square(between)/Mean Squar(within) != 1

c. **1 pts** Fill in the blanks for the degrees of freedom of the ANOVA F-test statistic: $F(2, 19)$

d. **1 pts** What is the p-value of the ANOVA F-test?

   p-value: 0.00447

e. **1 pts** According the the results of the ANOVA F-test, does light treatment affect phase shift? Use an α-level of 0.05.
   From the ANOVA table we got the p-value of 0.00447 which is smaller than α-level of 0.05. Hence we have to reject the null hypothesis and conclude that light treatment affect the phase shift.

Part B. Simple Linear Regression

We are going to use regression analysis to estimate the performance of CPUs based on the maximum number of channels in the CPU. This data set comes from the UCI Machine Learning Repository.

The data file includes the following columns:

- *vendor*: vendor of the CPU
- *chmax*: maximum channels in the CPU
- *performance*: published relative performance of the CPU

The data is in the file “machine.csv”. To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`.

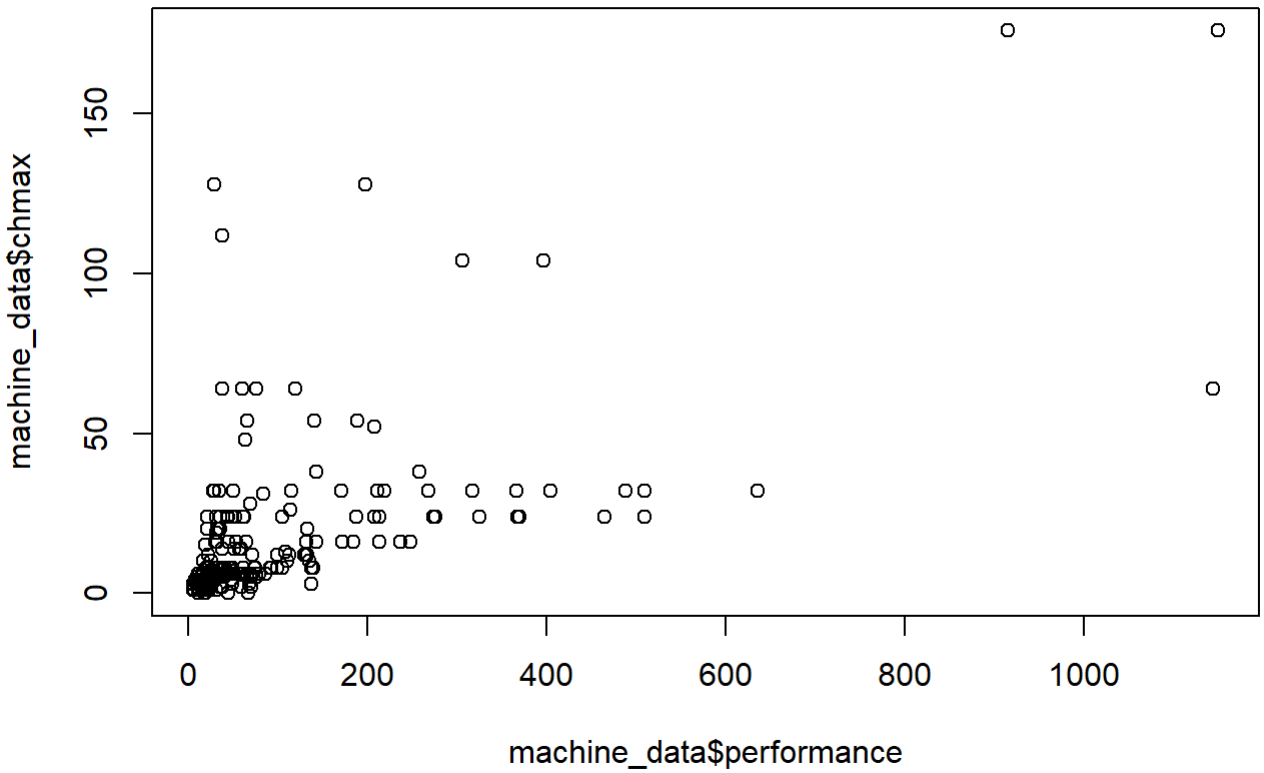
```
# Read in the data
machine_data = read.csv("machine.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(machine_data, 3)
```

```
##    vendor chmax performance
## 1 adviser   128         198
## 2  amdahl    32         269
## 3  amdahl    32         220
```

Question B1: Exploratory Data Analysis - 9 pts

a. **3 pts** Use a scatter plot to describe the relationship between CPU performance and the maximum number of channels. Describe the general trend (direction and form). Include plots and R-code used.

```
plot(machine_data$performance, machine_data$chmax)
```



From the above scatter plot we can say that there is some positive relationship between performance and number of channels. As the number of channel increased from 0 to 70 the

performance also increases. We can also say that there is no strong linear relationship between them. We can also see that there are some ouliers.To get a good idea of their relationship we need to check the correlation coefficient

b. **3 pts** What is the value of the correlation coefficient between *performance* and *chmax*? Please interpret the strength of the correlation based on the correlation coefficient.

```
cor(machine_data$performance, machine_data$chmax)
```

```
## [1] 0.6052093
```

From above correlation coefficient we can say that there is strong relationship between both variables.

c. **2 pts** Based on this exploratory analysis, would you recommend a simple linear regression model for the relationship? **Yes, based on the above analysis I would recommend a simple linear regression model.**

d. **1 pts** Based on the analysis above, would you pursue a transformation of the data? *Do not transform the data*. **Yes, I would transform the data and check if there is any improvement in the liner relationship between both variables.**

Question B2: Fitting the Simple Linear Regression Model - 11 pts

Fit a linear regression model, named *model1*, to evaluate the relationship between performance and the maximum number of channels. *Do not transform the data*. The function you should use in R is:

```
model1 = lm(performance ~ chmax, machine_data)
summary(model1)
```

```
##
## Call:
## lm(formula = performance ~ chmax, data = machine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -486.47  -42.20  -22.20   20.31  867.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2252    10.8587   3.428 0.000733 ***
## chmax         3.7441     0.3423  10.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128.3 on 207 degrees of freedom
## Multiple R-squared:  0.3663, Adjusted R-squared:  0.3632
## F-statistic: 119.6 on 1 and 207 DF,  p-value: < 2.2e-16
```

a. **3 pts** What are the model parameters and what are their estimates?
$\hat{\beta}_0$ = **37.2252**
$\hat{\beta}_1$ = **3.7441**
$\hat{\sigma}$ = **128.3**

b. **2 pts** Write down the estimated simple linear regression equation. **performance = 37.2252 + 3.7441(chmax)**

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem.
$\hat{\beta}_1$ = **3.7441** Which means that if we increase the number of channel by 1 the performance increases by 3.7441 unit.

d. **2 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level?

```
confint(model1,level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 15.817392 58.633048
## chmax        3.069251  4.418926
```

The β_1 statistically significant at 95% confidence interval as it does not contain 0 in the confidence interval range.

e. **2 pts** Is β_1 statistically significantly positive at an α-level of 0.01? What is the approximate p-value of this test?

```
tval = 10.938
1-pt(tval,207)
```

```
## [1] 0
```

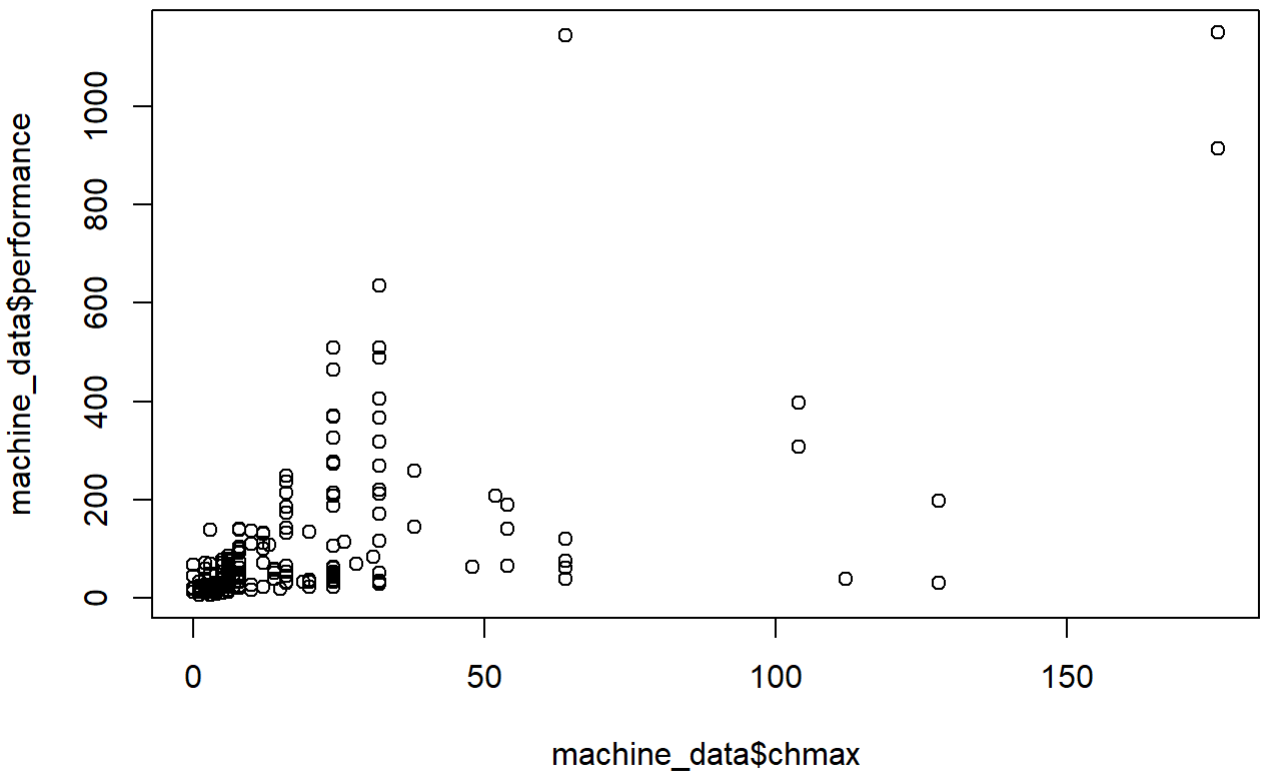
Here the p-value is very small(almost 0) than α-level of 0.01, hence we can say that β_1 statistically significantly positive. R is rounding up the p-value to 0 because its a very small number.

Question B3: Checking the Assumptions of the Model - 8 pts

Create and interpret the following graphs with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

a. **2 pts** Scatterplot of the data with *chmax* on the x-axis and *performance* on the y-axis

```
plot(machine_data$chmax, machine_data$performance)
```

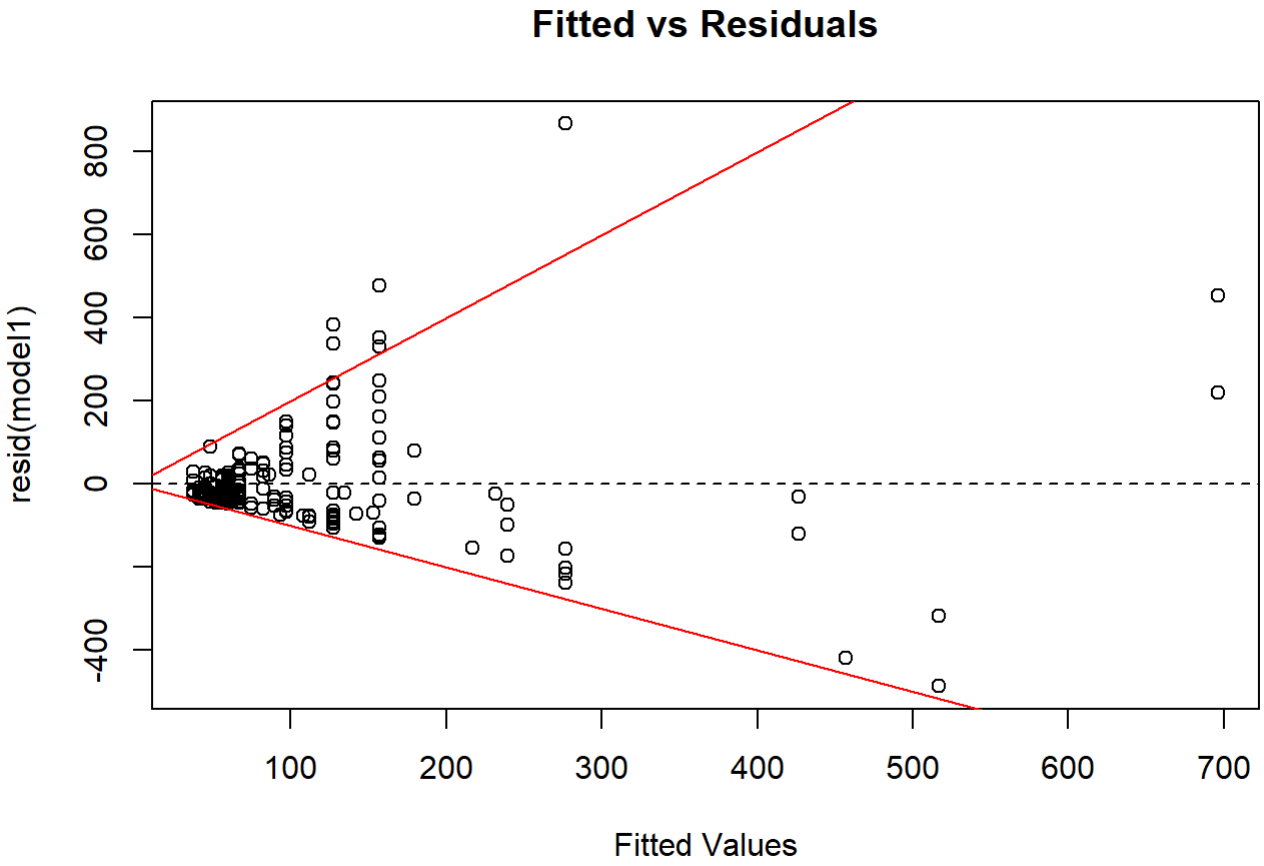


Model Assumption(s) it checks: The above scatter plot check the linearity assumption between "chmax" and "performance" variable.

Interpretation:
The above plot indicates that there is some positive linear relationship between the two variables. It also shows some out-liars on the top right corner of the graph.

b. **3 pts** Residual plot - a plot of the residuals, $\hat{\epsilon}_i$, versus the fitted values, $\hat{y}_i$

```
plot(fitted(model1),resid(model1),main="Fitted vs Residuals",
xlab="Fitted Values")
abline(h=0,lty=2)
abline(0,2, lty=1,col ="red")
abline(0,-1, lty=1,col= "red")
```

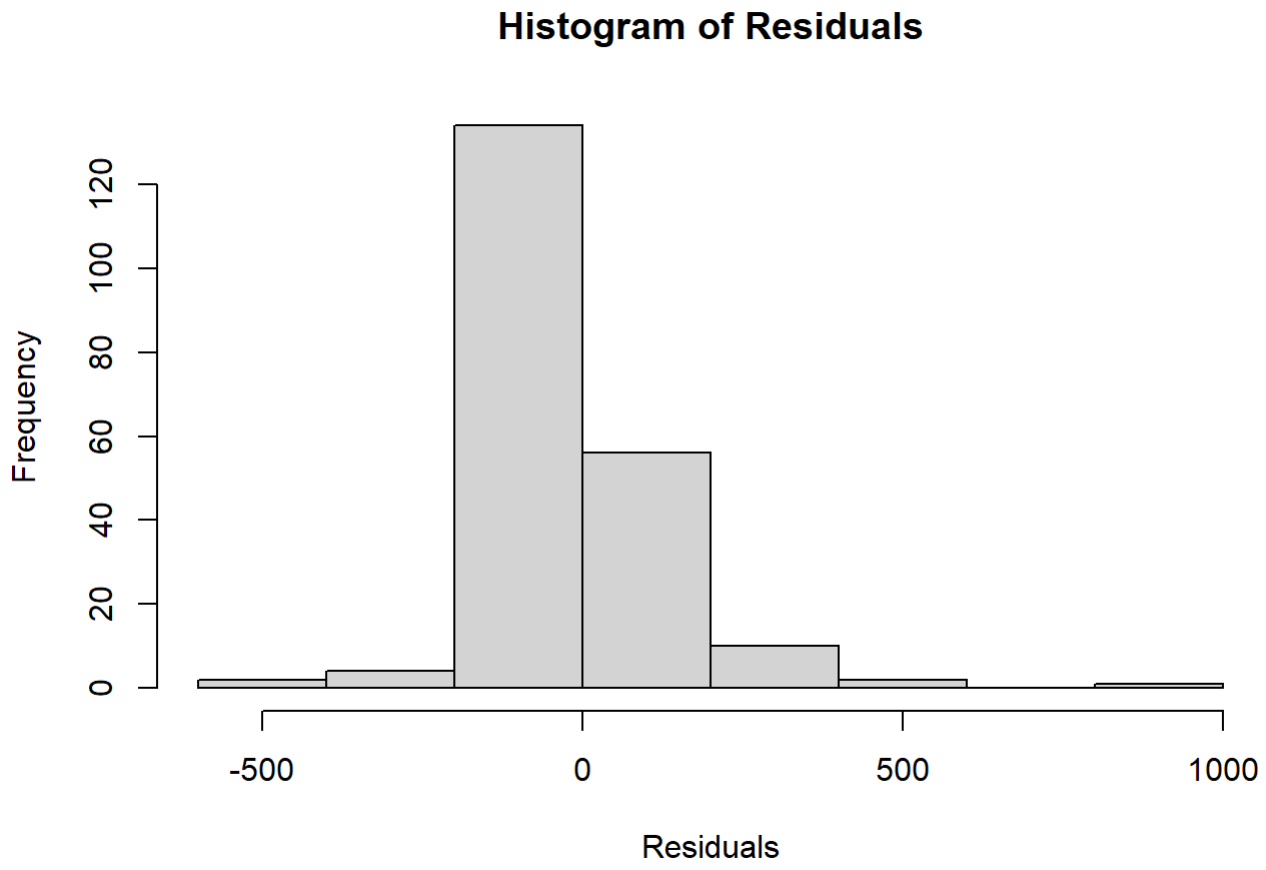


Model Assumption(s) it checks:

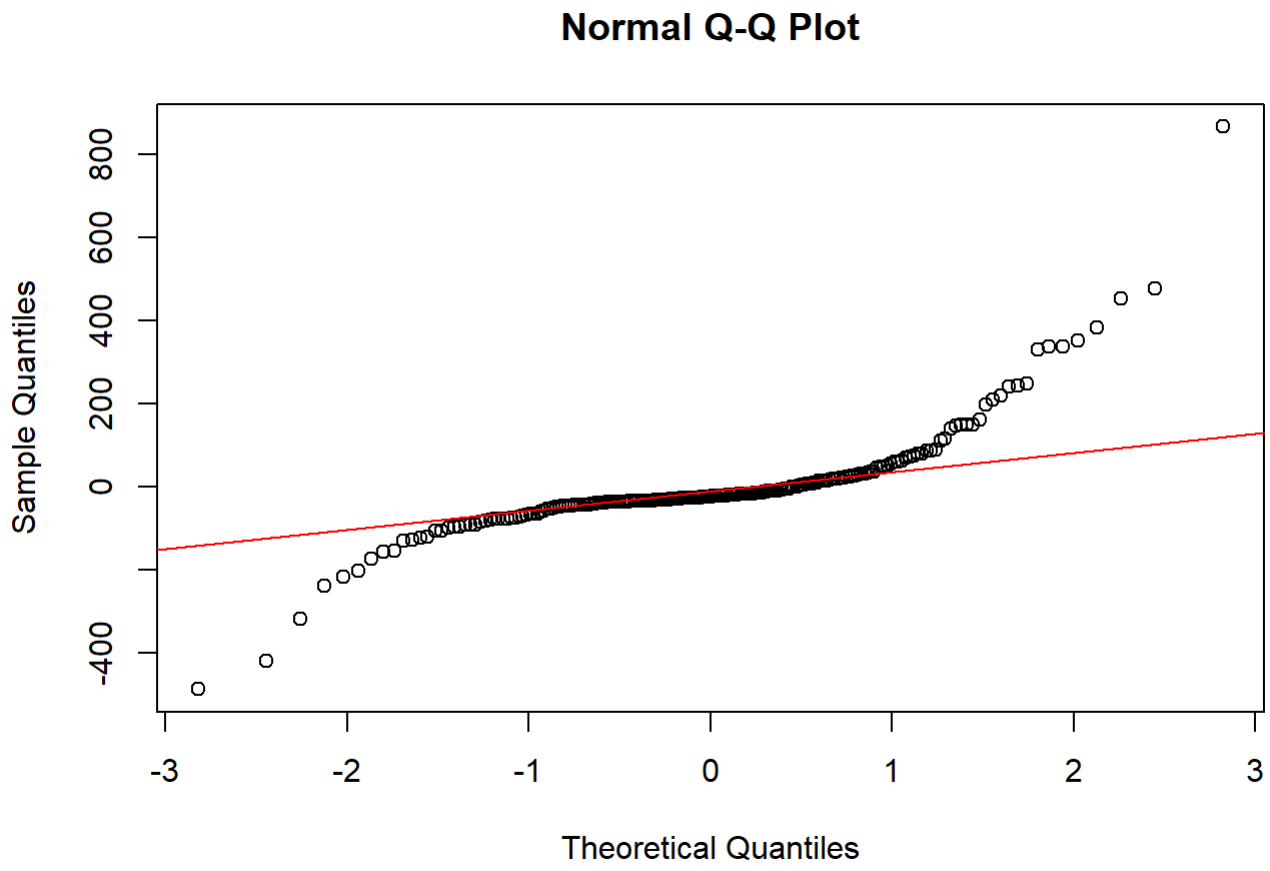
1. Homoscedasticity: The constant variance and the uncorrelated errors **Interpretation:**
   From the above graph we can say that residuals create megaphone effect and they are not randomly distributed around zero line. Hence, we can say that constant variance assumption and the assumption of uncorrelated errors doesn't hold.

c. **3 pts** Histogram and q-q plot of the residuals

```
hist(residuals(model1),main="Histogram of Residuals",
xlab="Residuals")
```




```
qqnorm(residuals(model1))
qqline(residuals(model1), col="red")
```



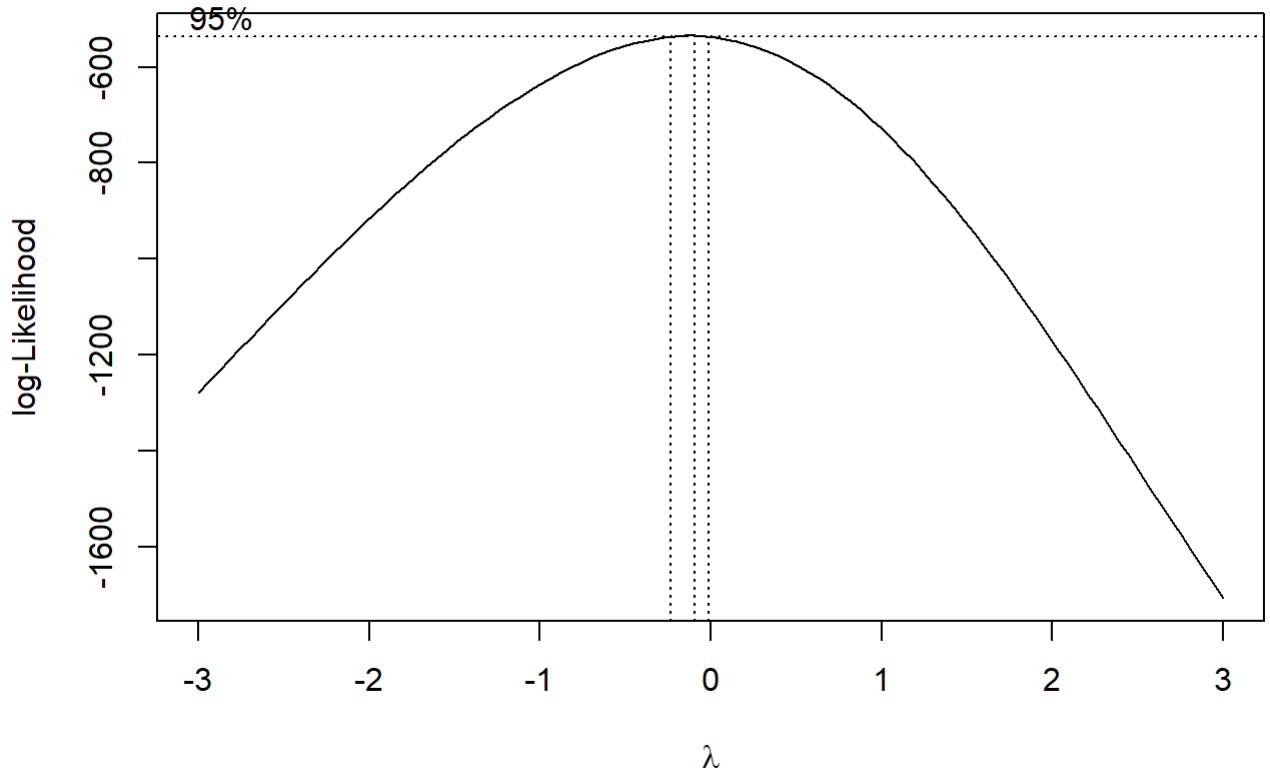
Model Assumption(s) it checks:
Normality Assumption

Interpretation:

From the above qq plot we can say that tails are extremely heavy at each end. Hence, the normal distribution doesn't hold. The histogram is also little skewed to right. ## Question B4: Improving the Fit - 10 pts

a. **2 pts** Use a Box-Cox transformation (`boxCox()`) to find the optimal λ value rounded to the nearest half integer. What transformation of the response, if any, does it suggest to perform?

```
library(MASS)
bc = boxcox(model1, lambda= seq(-3,3))
```




```
best.lam = bc$x[which(bc$y == max(bc$y))]
```

```
best.lam
```

```
## [1] -0.09090909
```

From the above graph and λ value we can use λ = 0 and perform simple log transformation.

b. **2 pts** Create a linear regression model, named *model2*, that uses the log transformed *performance* as the response, and the log transformed *chmax* as the predictor. Note: The variable *chmax* has a couple of zero values which will cause problems when taking the natural log. Please add one to the predictor before taking the natural log of it

```
machine_data$chmax <- if (any(machine_data$chmax==0)) {machine_data$chmax +1}
```

```
model2 <- lm(log(performance)~log(chmax), data = machine_data)
summary(model2)
```

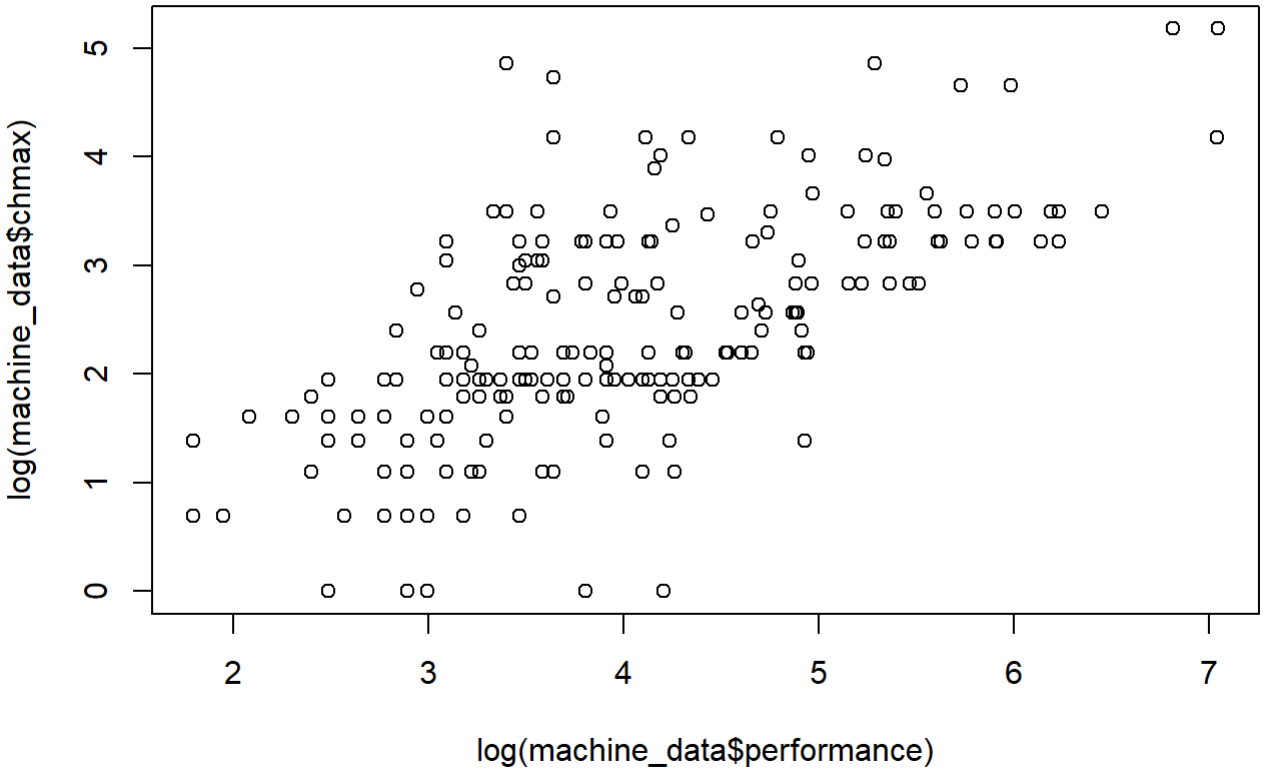
```
##
## Call:
## lm(formula = log(performance) ~ log(chmax), data = machine_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22543 -0.59429  0.01065  0.59287  1.85995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47655    0.14152    17.5   <2e-16 ***
## log(chmax)   0.64819    0.05401    12.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.807 on 207 degrees of freedom
## Multiple R-squared:  0.4103, Adjusted R-squared:  0.4074
## F-statistic:   144 on 1 and 207 DF,  p-value: < 2.2e-16
```

e. **2 pts** Compare the R-squared values of *model1* and *model2*. Did the transformation improve the explanatory power of the model?

***model1 gave the R-squared value of 0.3663 while model2 gave R-squared of 0.4103 which is little better than model1.**

c. **4 pts** Similar to Question B3, assess and interpret all model assumptions of *model2*. A model is considered a good fit if all assumptions hold. Based on your interpretation of the model assumptions, is *model2* a good fit?

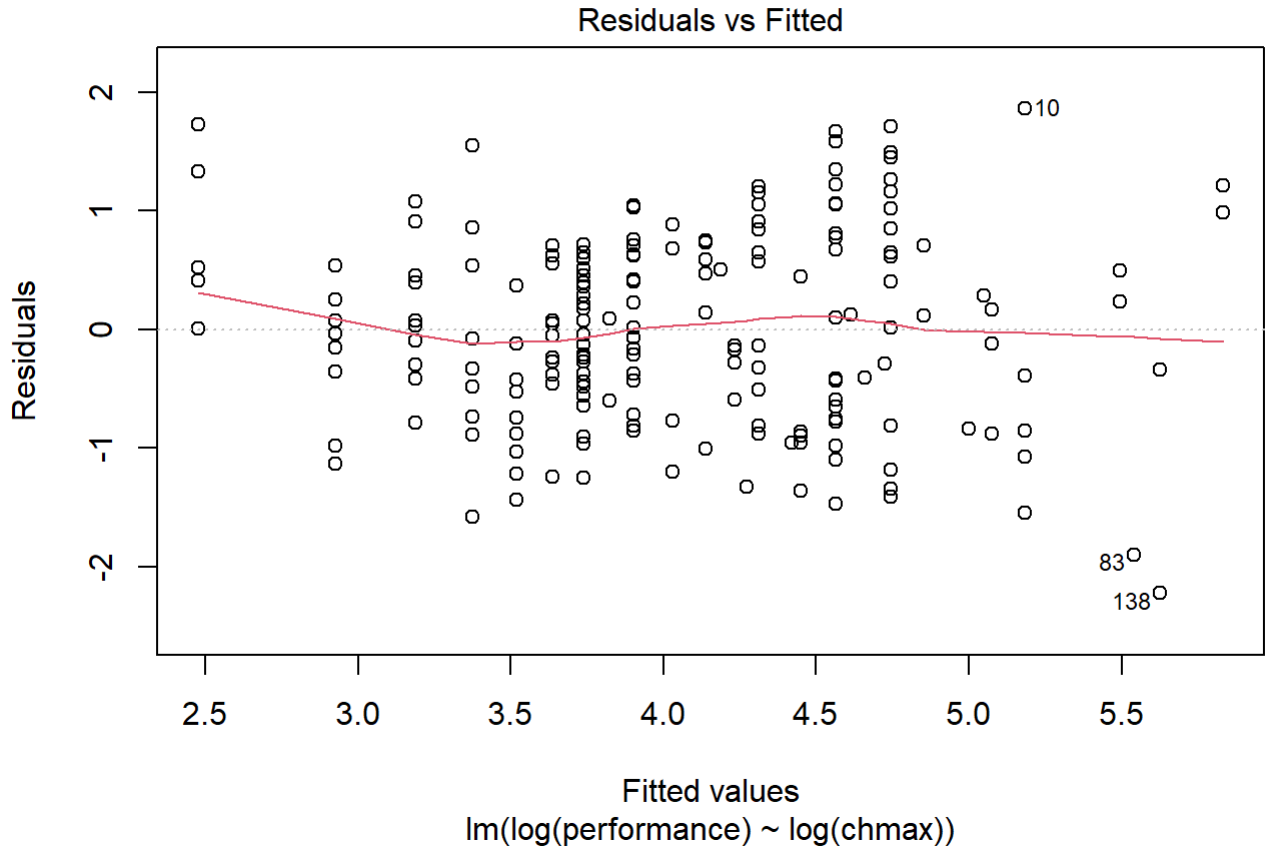
```
plot(log(machine_data$performance), log(machine_data$chmax))
```

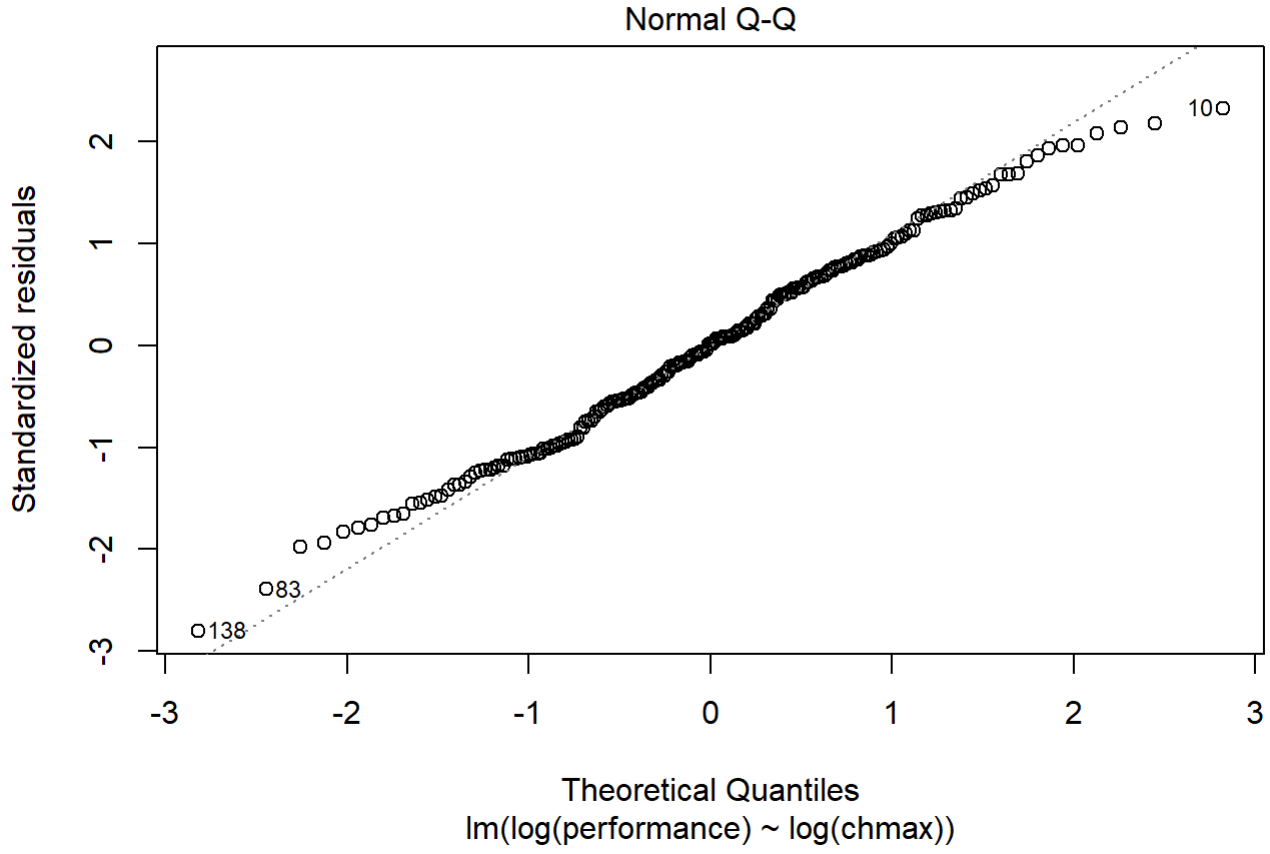


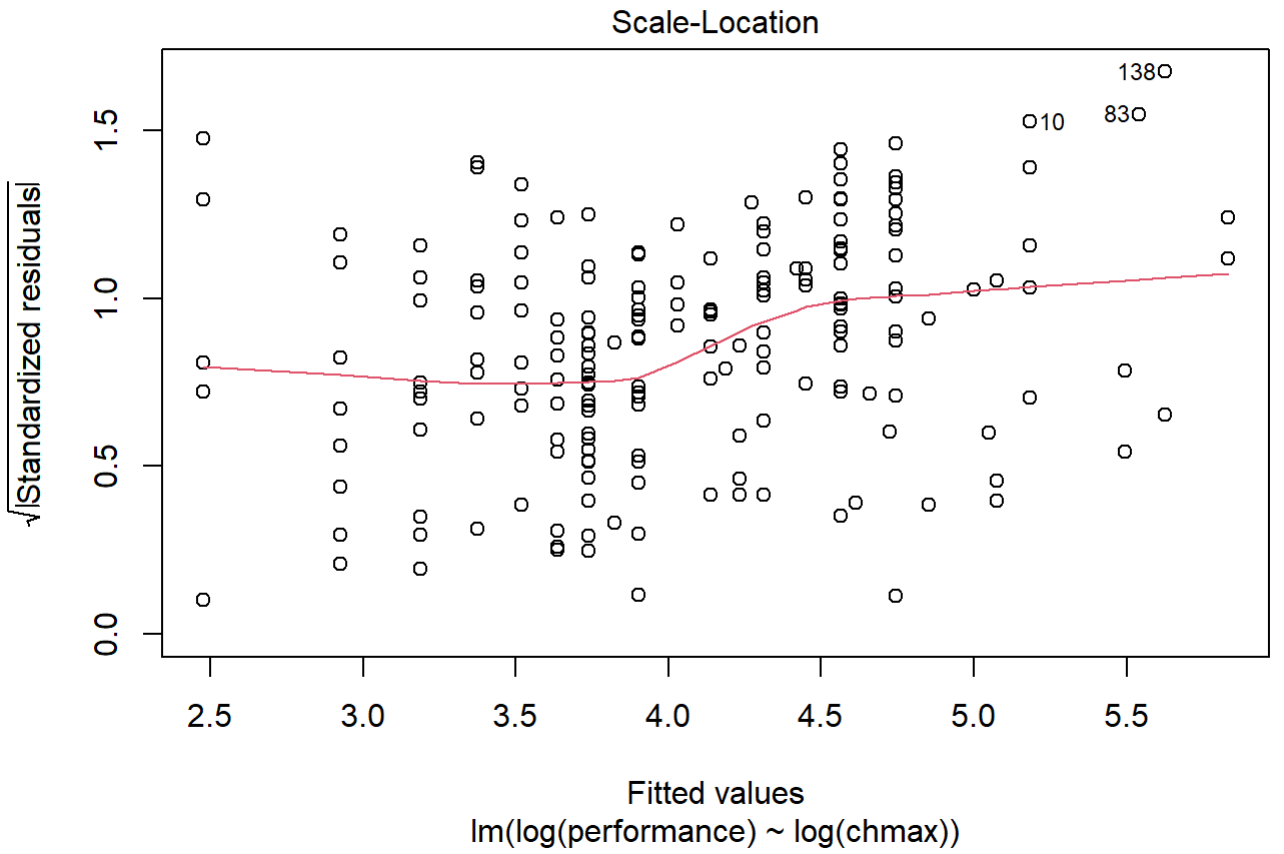
1)

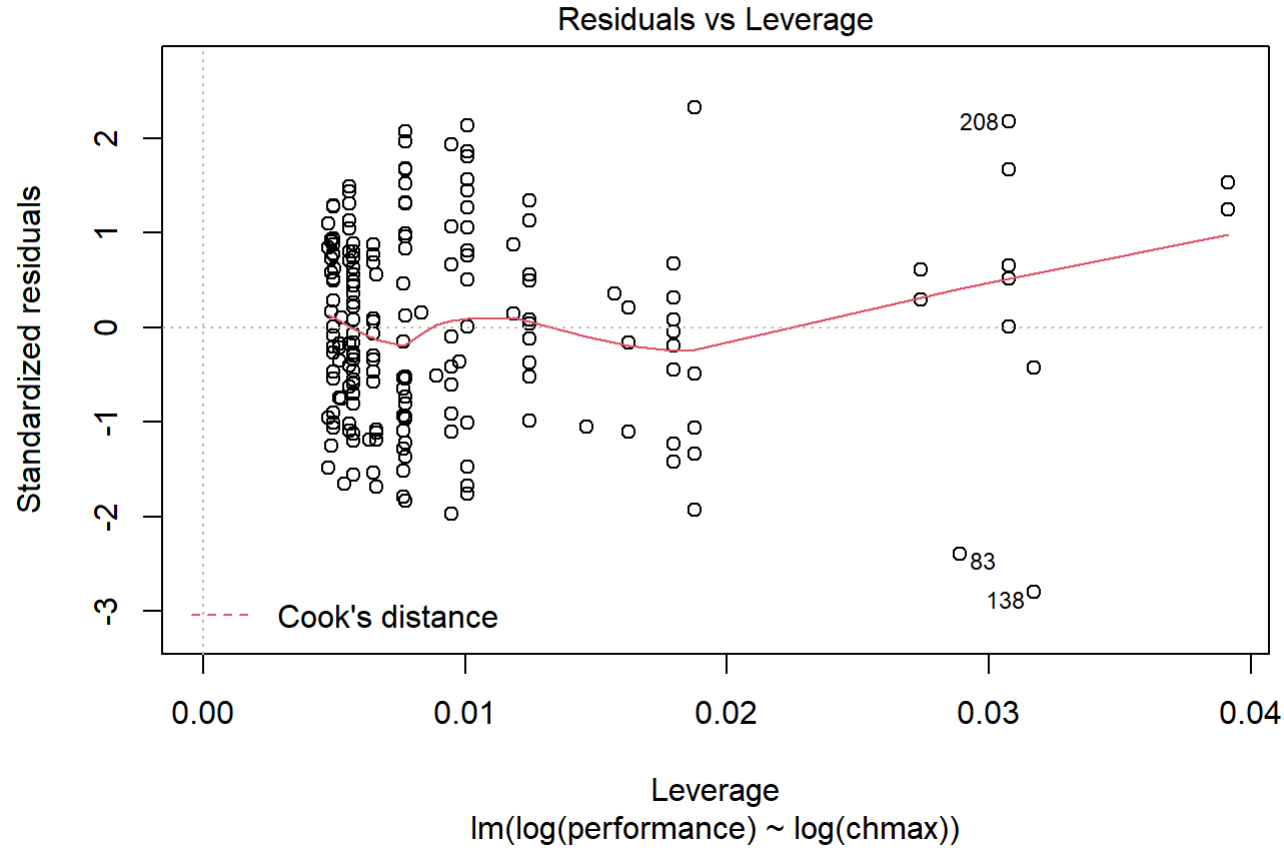
Linearity assumption
From the above graph we can say there is strong posotive linear relationship between log(performance) and log(chmax) variables.

```
plot(model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(performance) ~ log(chmax))
10
83
138


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(performance) ~ log(chmax))
10
83
138


Scale-Location
√|Standardized residuals|
138
83
10
0.0
0.5
1.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Fitted values
lm(log(performance) ~ log(chmax))


Residuals vs Leverage
Standardized residuals
208
83
138
Cook's distance
-3
-2
-1
0
1
2
0.00
0.01
0.02
0.03
0.04
Leverage
lm(log(performance) ~ log(chmax))

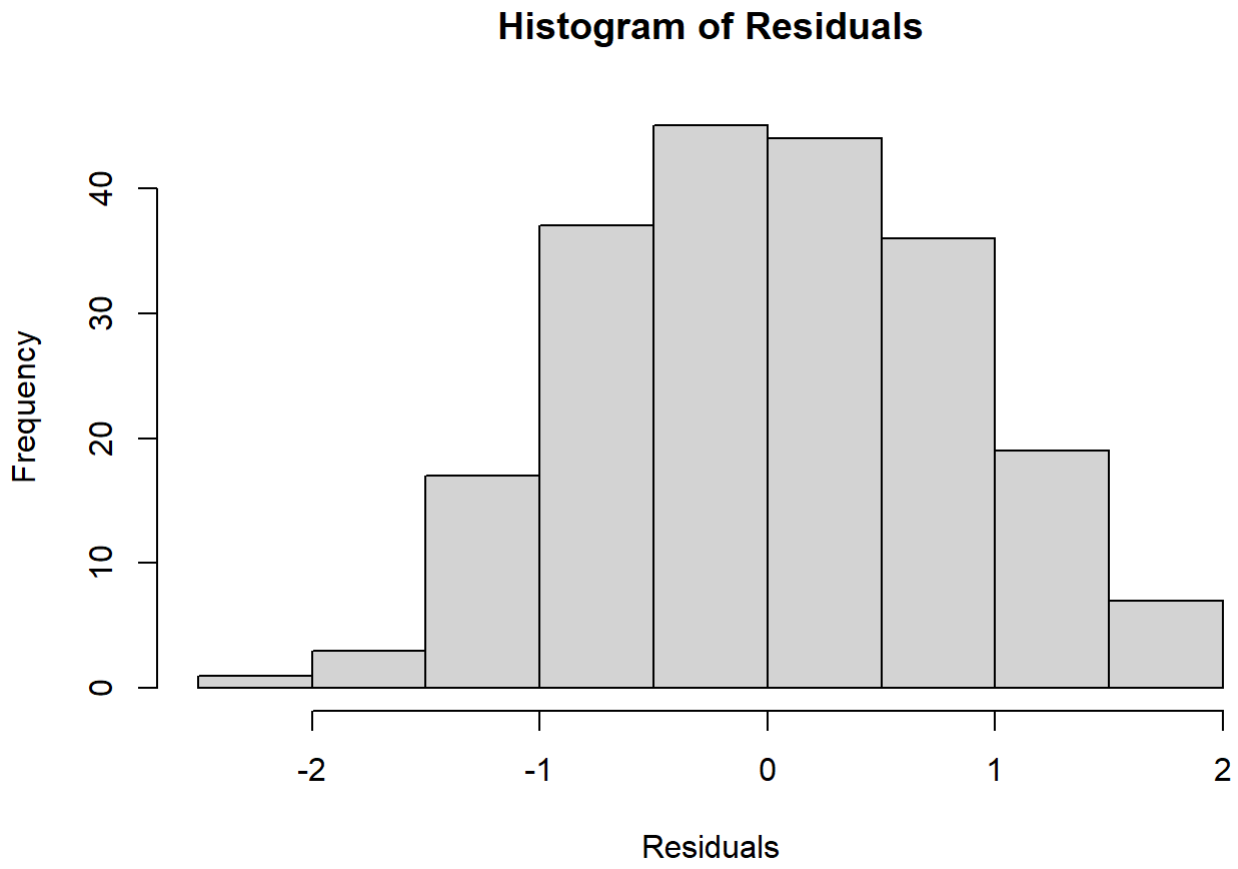
2) Constant variance and uncorrelated errors

The Residuals vs Fitted graph shows that residuals are randomly distributed around zero line which indicates that the model holds constant variance and uncorrelated errors assumptions.

3) Normality Assumption

The above qq plot indicates that almost all the points fall along the line. There is we small deviation at both ends but we can say the it holds the normality assumption. Below shown histogram also confirms that.

```
hist(residuals(model2),main="Histogram of Residuals",
xlab="Residuals")
```



From above analysis we can say that model2 holds all the assumption and is better fit than model1.

Question B5: Prediction - 3 pts

Suppose we are interested in predicting CPU performance when `chmax = 128`. Please make a prediction using both *model1* and *model2* and provide the 95% prediction interval of each prediction on the original scale of the response, *performance*. What observations can you make about the result in the context of the problem?

```
new1 = data.frame(chmax = 128)
predict(model1,new1,interval='prediction',level=.95)
```

```
##        fit      lwr      upr
## 1 516.4685 252.2519 780.6851
```

prediction interval on original scale

```
new2 = data.frame(chmax = 128)
exp(predict(model2,new2,interval='prediction',level=.95))
```

```
##        fit      lwr      upr
## 1 276.3256 54.90877 1390.594
```

When we compare the prediction intervals of the both the models we can say that model2 has very wide range. If we look at our data set we can clearly see that when "chmax" is > 100 the "performance" is between 0 to 1500 and hence our model is giving such a wide range for "chmax" =128.

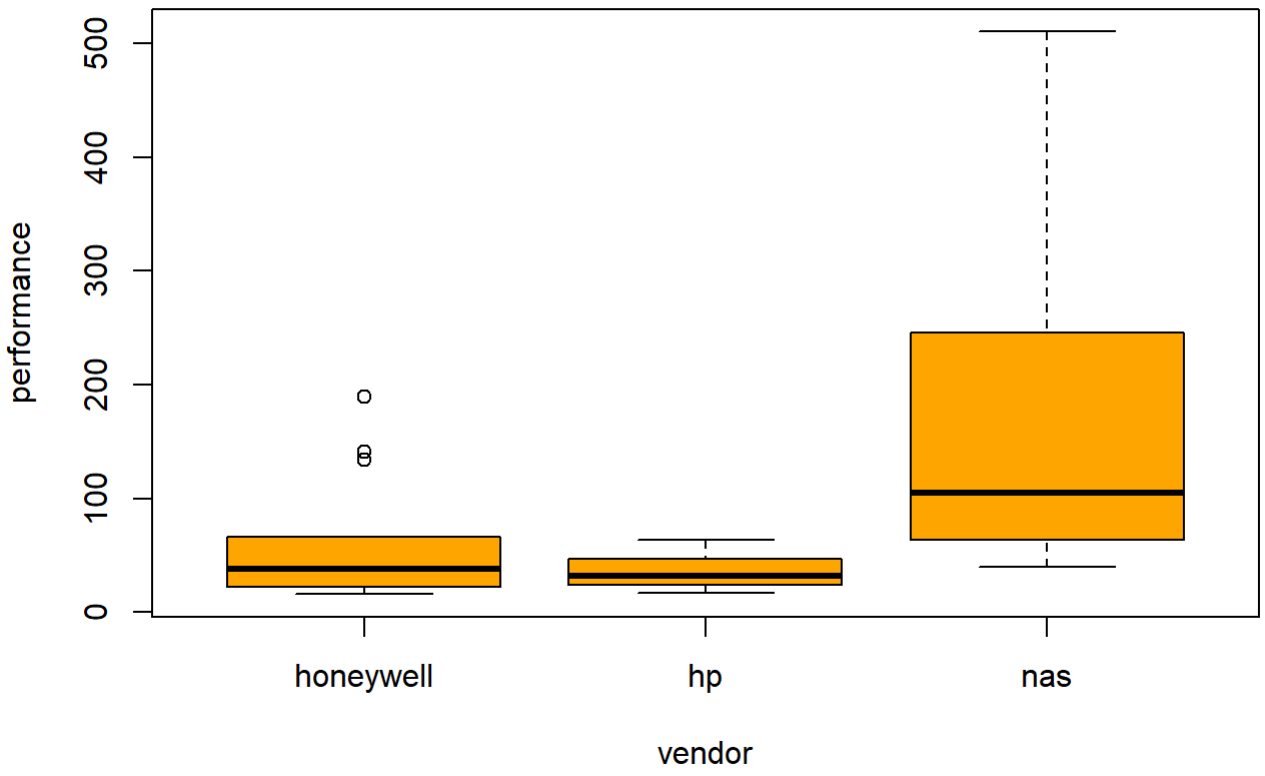
Part C. ANOVA - 8 pts

We are going to continue using the CPU data set to analyse various vendors in the data set. There are over 20 vendors in the data set. To simplify the task, we are going to limit our analysis to three vendors, specifically, honeywell, hp, and nas. The code to filter for those vendors is provided below.

```
# Filter for honeywell, hp, and nas
data2 =machine_data[machine_data$vendor %in% c("honeywell", "hp", "nas"), ]
data2$vendor = factor(data2$vendor)
```

1. **2 pts** Using `data2` , create a boxplot of *performance* and *vendor*, with *performance* on the vertical axis. Interpret the plots.

```
boxplot(performance~vendor, col= "orange", data = data2)
```



From the above plot we can say that all three have different means. The “nas” has higher mean than other two vendors. “nas” contains more data than other two vendors and also have more spread out data.

2. **3 pts** Perform an ANOVA F-test on the means of the three vendors. Using an α-level of 0.05, can we reject the null hypothesis that the means of the three vendors are equal? Please interpret.

```
anova_model = aov(performance~vendor, data = data2)
summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## vendor       2 154494   77247   6.027 0.00553 **
## Residuals   36 461443   12818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.tables(anova_model,type= "means")
```

```
## Tables of means
## Grand mean
##
## 112.8718
##
##  vendor
##     honeywell    hp   nas
##         60.46 36.43 176.9
## rep     13.00  7.00  19.0
```

from above test we can reject the null hypothesis of equal means for α-level of 0.05 as the p-value is less than 0.05.

3. **3 pts** Perform a Tukey pairwise comparison between the three vendors. Using an α-level of 0.05, which means are statistically significantly different from each other?

```
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = performance ~ vendor, data = data2)
##
## $vendor
##                     diff        lwr      upr     p adj
## hp-honeywell  -24.03297 -153.76761 105.7017 0.8934786
## nas-honeywell 116.43320   16.82659 216.0398 0.0188830
## nas-hp        140.46617   18.11095 262.8214 0.0214092
```

From the above test we can say that nas-honeywel and nas-hp has statistically significantly different means from each other as they do not contain any zero in their lower and upper rang and also they have their p=value is less than 0.05.
FOr “Hp-honeywell”, the p-value is higher than 0.05 that is why we can not reject the null hypothesis for that pair.