HW1 Peer Assessment

Part A. ANOVA

Additional Material: ANOVA tutorial

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

Consider the following incomplete R output:

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	2	7.2244	3.6122	7.2896	0.004
Error	19	9.415	0.4955		
TOTAL	21	16.6394			

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.

1: Sample mean of phase shift hours in melatonin production without treatment = -0.3088

2: Sample mean of phase shift hours in melatonin production with treatment to knees = -0.3357

3: Sample mean of phase shift hours in melatonin production with treatment to eyes = -1.5514

Question A3 - 5 pts

Use the ANOVA table in Question A1 to write the:

a. **1 pts** Write the null hypothesis of the ANOVA F-test, H_0

Ho: 1 = 2 = 3

b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A

Ha: 1 2 or 1 3 or 2 3

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

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

P-value = 0.004

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.

P-value is less than α-level of 0.05, therefore we reject the null hypothesis and conclude the difference of phase shift hours among treatment and control groups are not statistically significant

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

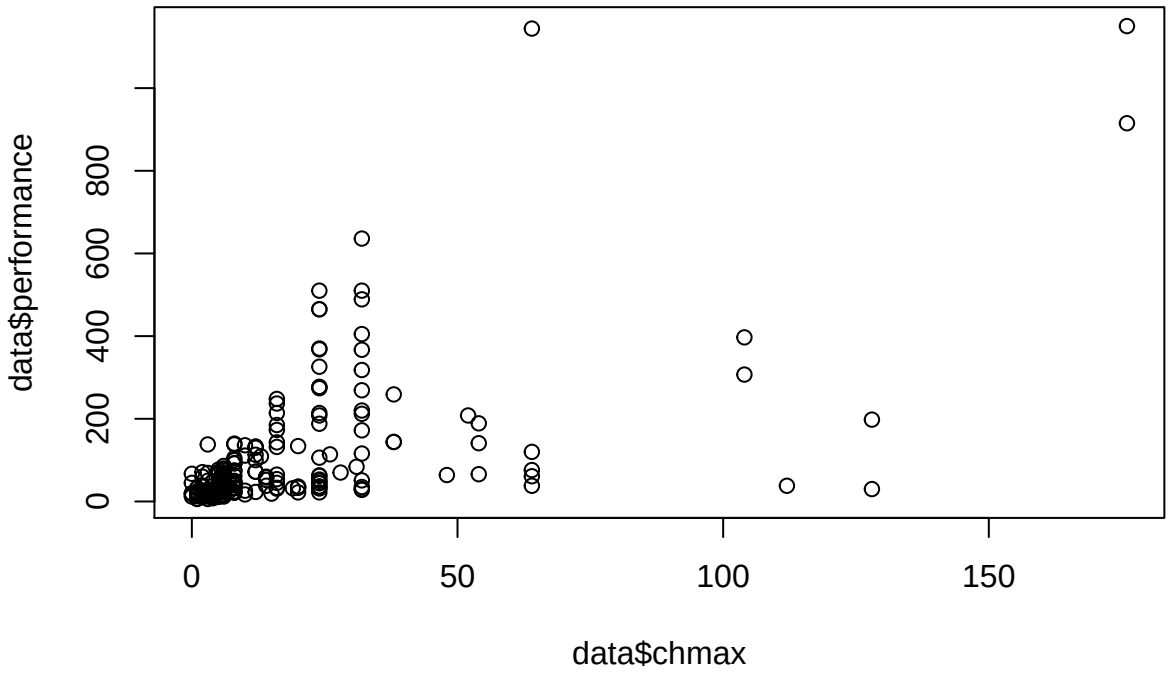
The data is in the file "machine.csv". To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`.

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

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.

```
# Your code here...
plot(data$chmax, data$performance)
```



The scatter plot above suggests there's positive relationship between CPU performance and max channels except for a few outliers, and the form appears to be curvature.

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.

```
# Your code here...
cor(data$chmax, data$performance)
```

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

The correlation coefficient of 0.6052 suggests a moderate positive relationship between performance and chmax.

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

Since the linear relationship is not strong, I wouldn't recommend a simple linear regression.

d. **1 pts** Based on the analysis above, would you pursue a transformation of the data?

I would pursue a transformation to improve nonlinearity.

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:

```
# Your code here...
model1 = lm(performance ~ chmax, data)
summary(model1)
```

```
##
## Call:
## lm(formula = performance ~ chmax, data = 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?

$\hat{\beta}_0$ = Estimated incercept parameter (estimated expected value of the response variable, when the predicting variable equals zero) = 37.2252

$\hat{\beta}_1$ = Estimated slope parameter (estimated expected change in the response variable associated with one unit of change in the predicting variable) = 3.7441

b. **2 pts** Write down the equation for the simple linear regression model.

performance = 37.2252 + 3.7441 * chmax

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem.

As number of channel increase by 1 unit, the expected unit increase in performance is 3.7441

d. **2 pts** Find a 95% confidence interval for the $\hat{\beta}_1$ parameter. Is $\hat{\beta}_1$ statistically significant at this level?

```
# Your code here...
confint(model1)
```

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

Confidence interval for $\hat{\beta}_1$ is (3.0693, 4.4189). $\hat{\beta}_1$ is statistically significant at this level.

e. **2 pts** Is $\hat{\beta}_1$ statistically significantly positive at an α-level of 0.01? What is the approximate p-value of this test?

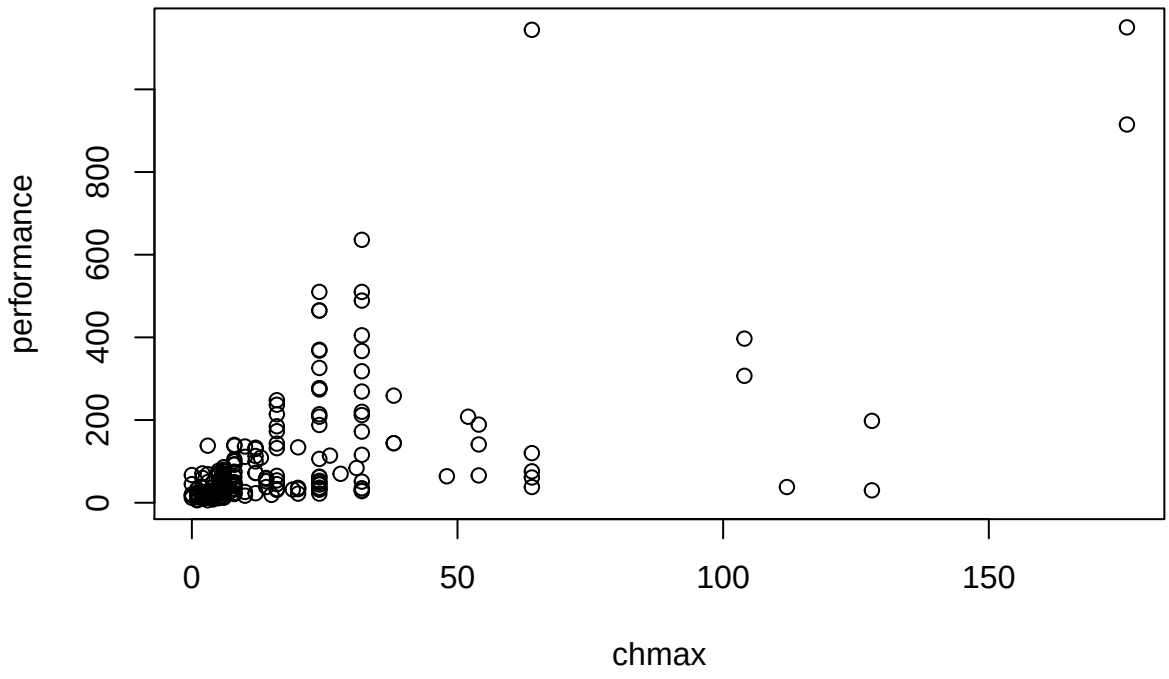
$\hat{\beta}_1$ is statistically significantly positive at α-level of 0.01. P-value is <2e-16 0

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

```
# Your code here...
plot(data$chmax, xlab='chmax', data$performance, ylab='performance')
```



Model Assumption(s) it checks:

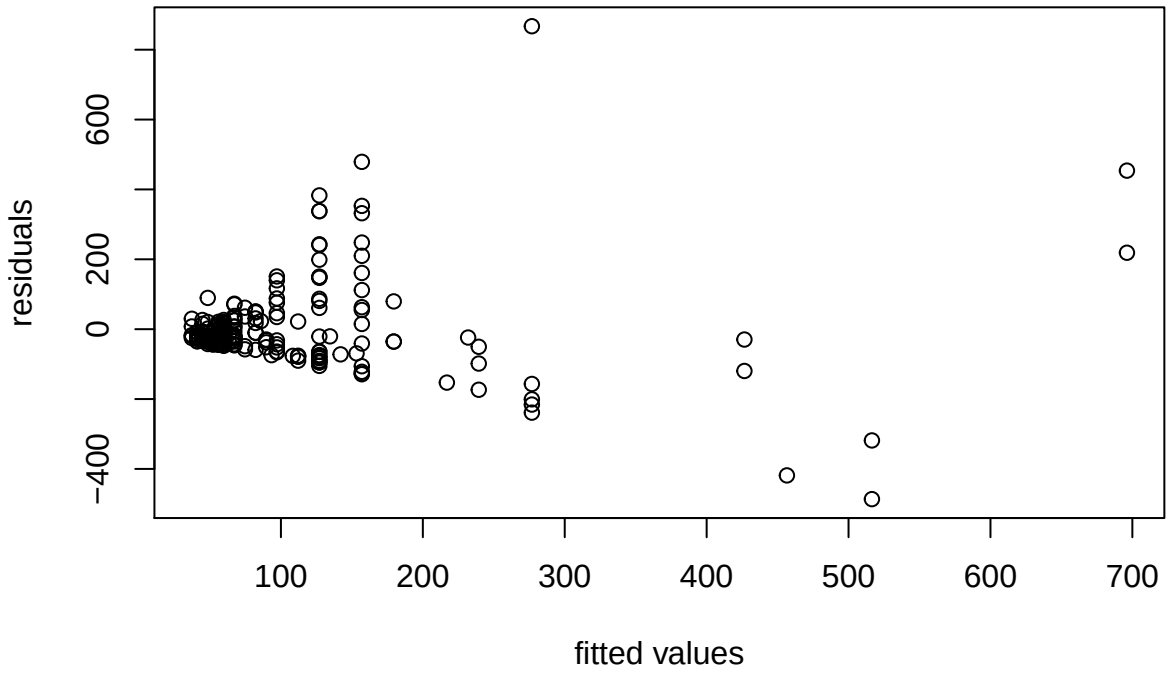
Linearity

Interpretation:

Curvature, therefore linearity assumption does not hold

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

```
# Your code here...
plot(model1$fitted.values,xlab='fitted values', model1$residuals,ylab='residuals')
```



Model Assumption(s) it checks:

Constant variance, Linearity, Independence

Interpretation:

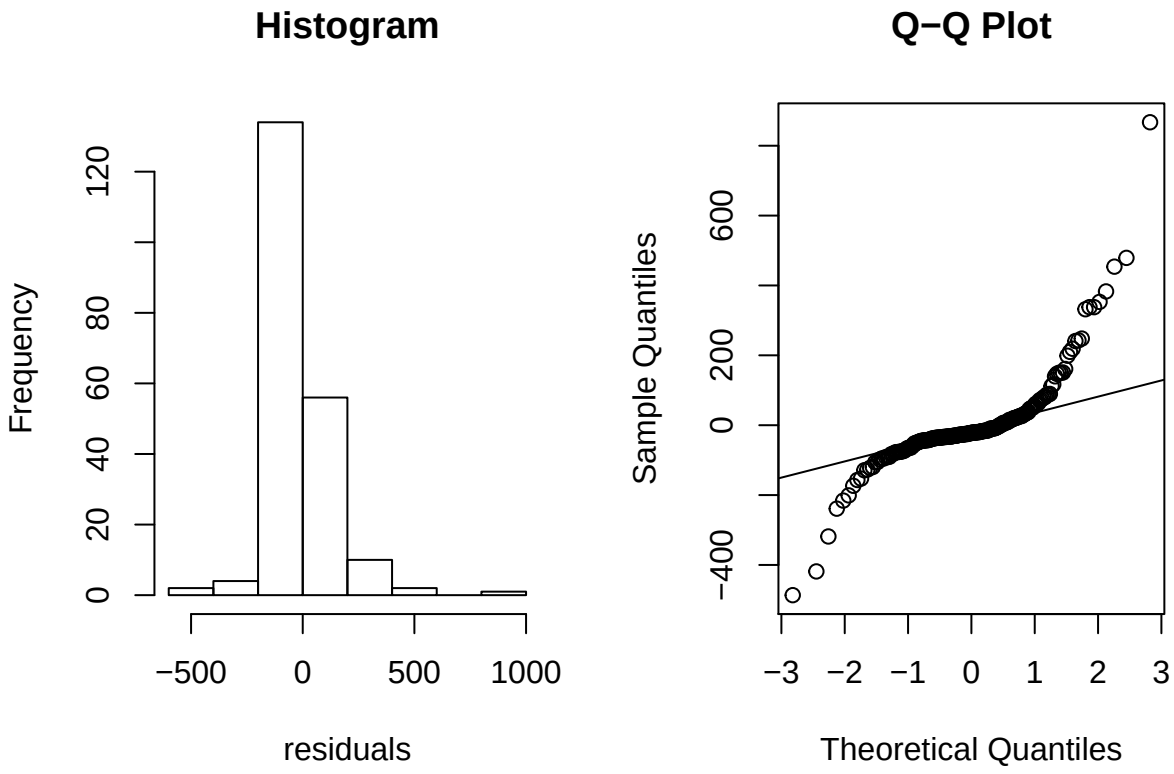
Residuals increase with increasing fitted value, therefore constant variance assumption does not hold

Curvature, therefore linearity assumption does not hold

Clustering of residuals on the left side, so the assumption of independence may not hold

c. **3 pts** Histogram and q-q plot

```r
# Your code here...
par(mfrow=c(1,2))
hist(model1$residuals, xlab = 'residuals', main = 'Histogram')
qqnorm(model1$residuals, main = 'Q-Q Plot')
qqline(model1$residuals)
```



Model Assumption(s) it checks:

Normality

Interpretation:

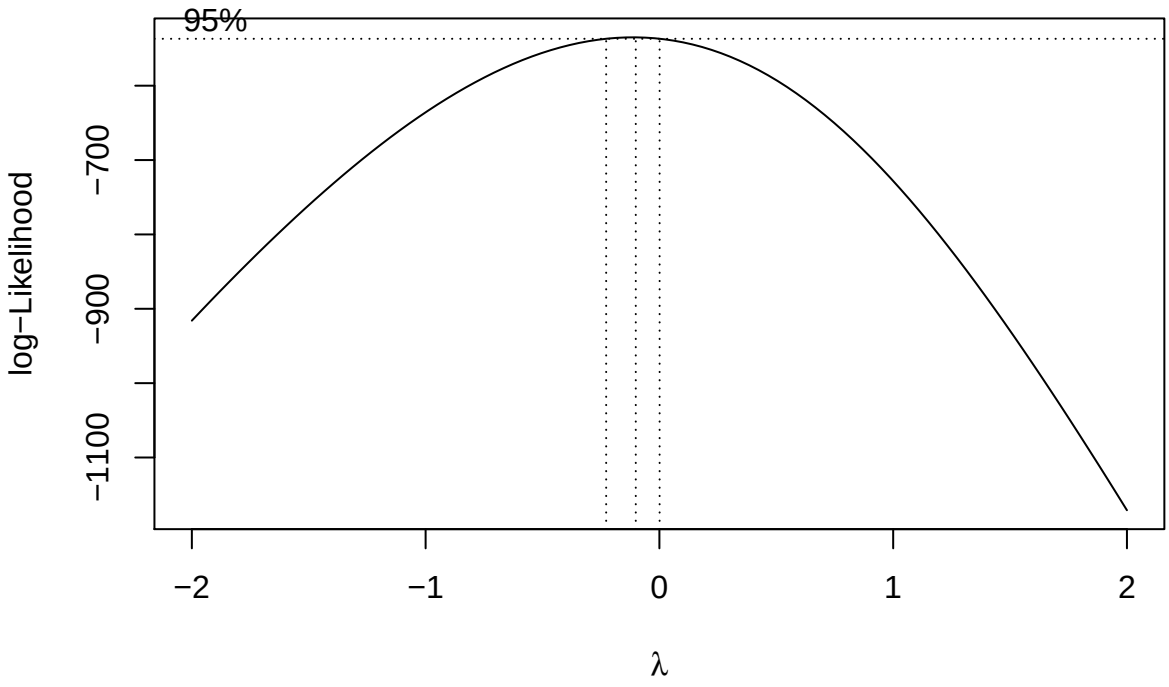
Histogram: Right-skewed, normality assumption does not hold

Q-Q: Heavy-tailed, normality assumption does not hold

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?

```r
# Your code here...
library(MASS)
boxcox(model1)
```



λ value of 0 suggests a normal logarithmic 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.

```
# Your code here...
data2 = data
data2$performance = log(data2$performance)
data2$chmax = log(data2$chmax+1)
model2 = lm(performance ~ chmax, data2)
summary(model2)
```

```
##
## Call:
## lm(formula = performance ~ chmax, data = data2)
##
## 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 ***
## chmax        0.64819    0.05401    12.0   <2e-16 ***
## ---
```

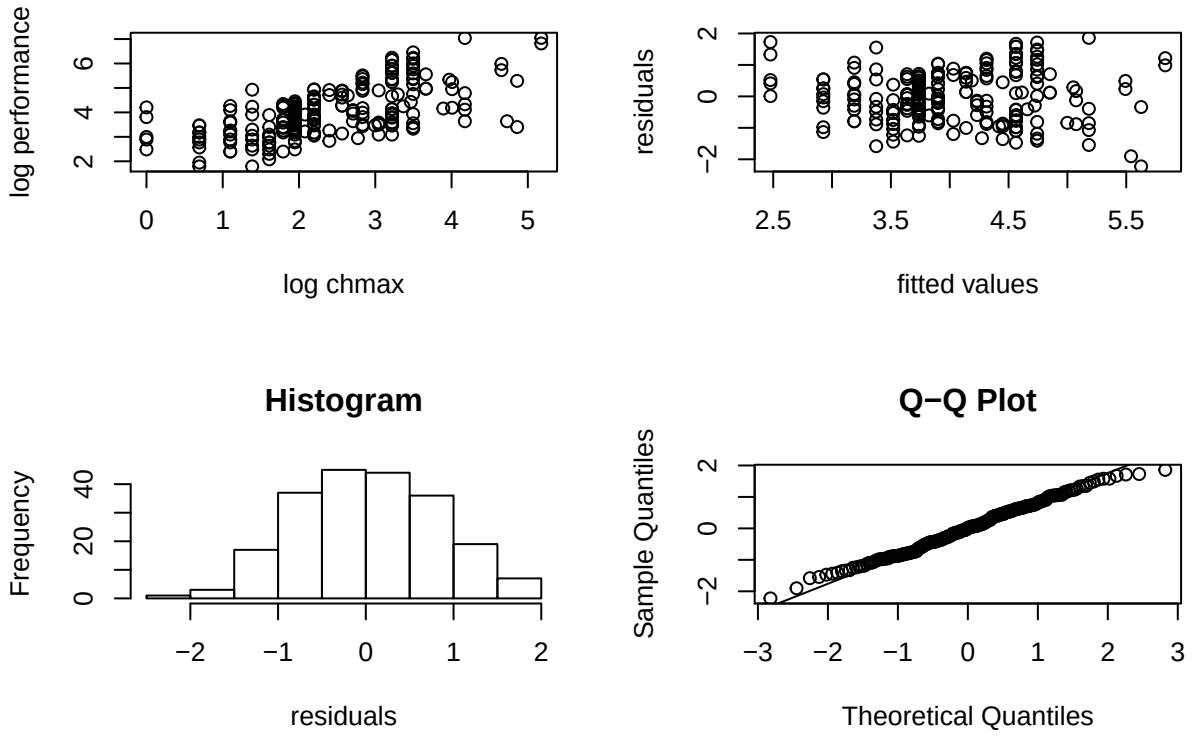
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.807 on 207 degrees of freedom
## Multiple R-squared:  0.4103, Adjusted R-squared:  0.4074
## F-statistic:   144 on 1 and 207 DF,  p-value: < 2.2e-16
```

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

R-squared value of model2 is around 0.41 which is higher than model1 (0.36), therefore the transformation appears to improve the explanatory power of the model.

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?

```
# Your code here...
par(mfrow=c(2,2))
plot(data2$chmax, xlab='log chmax', data2$performance, ylab='log performance')
plot(model2$fitted.values,xlab='fitted values', model2$residuals,ylab='residuals')
hist(model2$residuals, xlab = 'residuals', main = 'Histogram')
qqnorm(model2$residuals, main = 'Q-Q Plot')
qqline(model2$residuals)
```



1. Scatterplot of log performance vs log chmax assesses linearity, and the plot suggests a positive linear relationship, therefore linearity assumption holds.

2. Residuals vs fitted plot assesses linearity and constant variance. Random pattern around zero line suggests linearity assumption holds, and equal distribution suggests constant variance assumption holds.

3. Histogram plot assesses normality. The bell curve suggests normality assumption holds.

4. Q-Q plot assesses normality. Quantiles of residuals line up with normal quantiles on a straight line, therefore normality assumption holds.

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?

```
# Your code here...
new = data.frame(chmax = 128)
new2 = data.frame(chmax = log(128))
predict(model1, new, interval = 'prediction', level = 0.95)
```

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

```
exp(predict(model2, new2, interval = 'prediction', level = 0.95))
```

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

The predicted CPU performance for model2 is lower than model1, and the prediciton interval for model2 appears to be greater than model1.

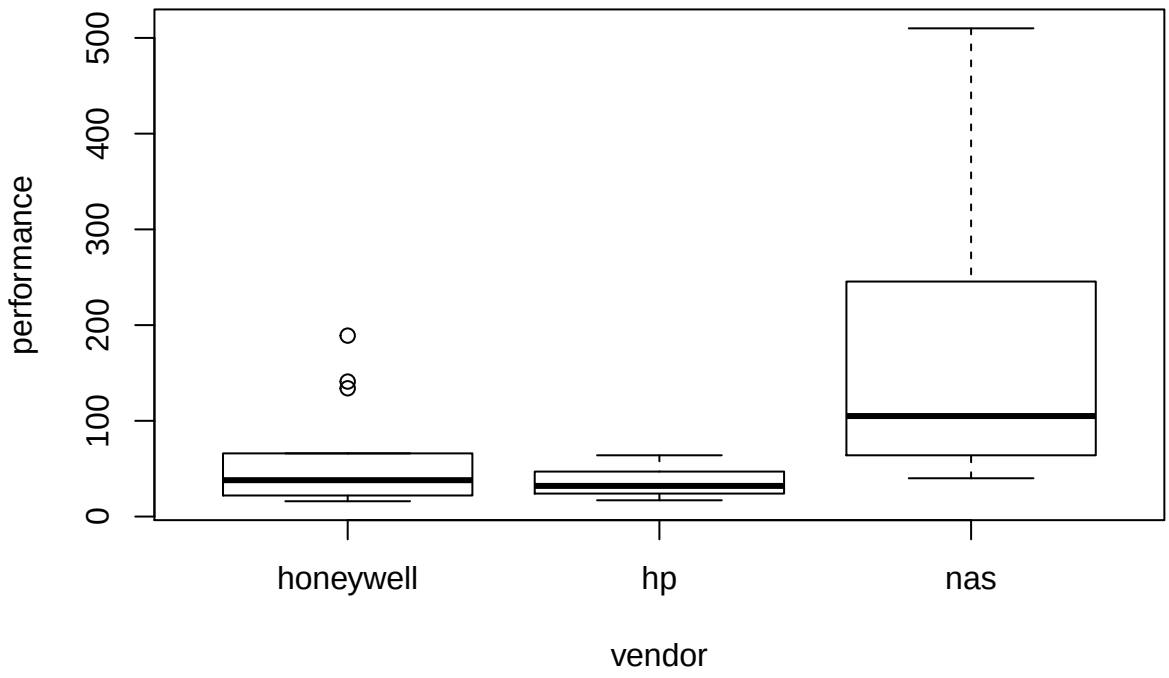
Part C. ANOVA - 8 pts

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 = data[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.

```
# Your code here...
boxplot(data2$performance~data2$vendor, xlab='vendor', ylab='performance')
```



There are differences in the means and between-variability among vendors, with mean performance of nas being the highest. There are also presence of differences in the within-variability within each vendor.

2. **3 pts** Perform an ANOVA F-test on the means of the three vendors. Using an α-level of 0.01, does the null hypothesis hold? Please interpret.

```
# Your code here...
model_aov = aov(performance~vendor, data2)
summary(model_aov)
```

```
##             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
```

P-value of 0.00553 is less than α-level of 0.01, therefore we reject the null hypothesis of equal means.

3. **3 pts** Perform a Tukey pairwise comparison between the three vendors (`TukeyHSD()`). Using an α-level of 0.01, which means are statistically significantly different from each other?

```
# Your code here...
TukeyHSD(model_aov)
```

```
##   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
```

The means of all pairs are not statistically significantly different among each other with α-level of 0.01 since the p-values are greater than α-level.