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

Creating data

```
treatment <- c('Control','Control','Control','Control','Control','Control','Control','Control',
               'Knees','Knees','Knees','Knees','Knees','Knees','Knees',
               'Eyes','Eyes','Eyes','Eyes','Eyes','Eyes','Eyes')
shift <- c(0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27,
           0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61,
           -0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83)
df <- data.frame(treatment, shift)
print(df)
```

```
##    treatment shift
## 1    Control  0.53
## 2    Control  0.36
## 3    Control  0.20
## 4    Control -0.37
## 5    Control -0.60
## 6    Control -0.64
## 7    Control -0.68
## 8    Control -1.27
## 9      Knees  0.73
## 10     Knees  0.31
## 11     Knees  0.03
## 12     Knees -0.29
## 13     Knees -0.56
## 14     Knees -0.96
## 15     Knees -1.61
## 16      Eyes -0.78
## 17      Eyes -0.86
## 18      Eyes -1.35
## 19      Eyes -1.48
## 20      Eyes -1.52
## 21      Eyes -2.04
## 22      Eyes -2.83
```

Question A1 - 3 pts

Consider the following incomplete R output:

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	?	?	3.6122	?	0.004
Error	?	9.415	?		
TOTAL	?	?			

Fill in the missing values in the analysis of the variance table.Note: Missing values can be calculated using the corresponding formulas provided in the lectures, or you can build the data frame in R and generate the ANOVA table using the aov() function. Either approach will be accepted.

```
model0 = aov(shift ~ treatment,data = df)
summary(model0)
```

```
##             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
```

```
model.tables(model0,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
```

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			

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 (i.e. explain what μ_1, μ_2, and μ_3 mean in words in the context of this problem). Find the estimates of these parameters.

```
model.tables(model0,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
```

μ_1: Average phase shift (measured in hours) in each subject’s daily cycle of melatonin production for subjects in the control group (no light) Value: -0.3087 hours

μ_2: Average phase shift (measured in hours) in each subject’s daily cycle of melatonin production for subjects in the eyes group Value: -1.551 hours

μ_3 Average phase shift (measured in hours) in each subject’s daily cycle of melatonin production for subjects in the knees group Value: -0.3357 hours

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: All three groups (2 treatments + control) will have the same mean effect on phase shift
b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A H_1: At least 1 of the three groups (2 treatments + control) will have a different mean effect on phase shift
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 = .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. Yes

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)
```

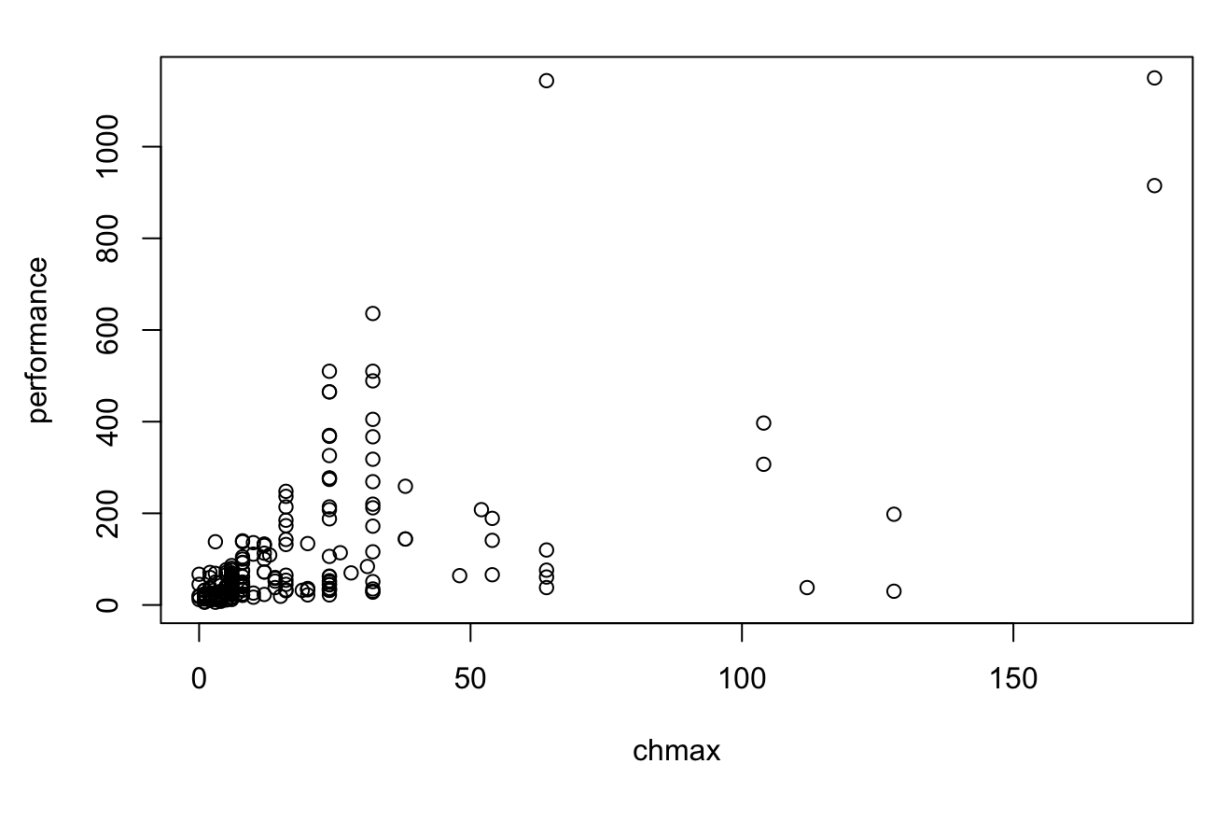
	vendor	chmax	performance
	<chr>	<int>	<int>
1	adviser	128	198
2	amdahl	32	269
3	amdahl	32	220

3 rows

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.

```
attach(data)
plot(chmax,performance)
```



There seems to be a positive relationship between performance and max # of channels, though there are some outliers and/or leverage points that have chmax much higher than the bulk of the group. It also appears the data is a “cone” shape so the data is not linear, meaning that chmax alone does not explain all of the variability in performance.

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...
model = lm(performance ~ chmax)
summary(model)
```

```
##
## Call:
## lm(formula = performance ~ chmax)
##
## 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
```

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

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

The correlation coefficient is 0.6052. This is a fairly strong positive correlation.

c. **2 pts** Based on this exploratory analysis, would you recommend a simple linear regression model for the relationship? I would say no. The coefficient of determination (r-squared) is .3663, indicating that only one third of the variability in performance can be explained by variability in chmax, which is not that good if it will be used as the sole predictor.
d. **1 pts** Based on the analysis above, would you pursue a transformation of the data? *Do not transform the data.* Yes, the funnel shape and the clustering of points at the lower end of chmax suggests to me that a logarithmic transformation may compress the data and lead to a more linear relationshipl

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?
   The model parameters are chmax and the intercept. The estimate of chmax is 3.7441 and the estimate of the intercept is 37.2252
b. **2 pts** Write down the estimated simple linear regression equation. y = 3.7441*chmax + 37.2252 + e
c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem. β_1, which is chmax in this case, is 3.7441. This suggests that for every additional maximum channel in the CPU, the performance metric will imporve by 3.7441
d. **2 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level?

```
confint(model1,level=.95)
```

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

The 95% confidence interval for β_1 is (3.069251,4.418926). That makes β_1 statistically significant (non zero) at that level.

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?

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

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

```
#confint(model1,level=.95)
```

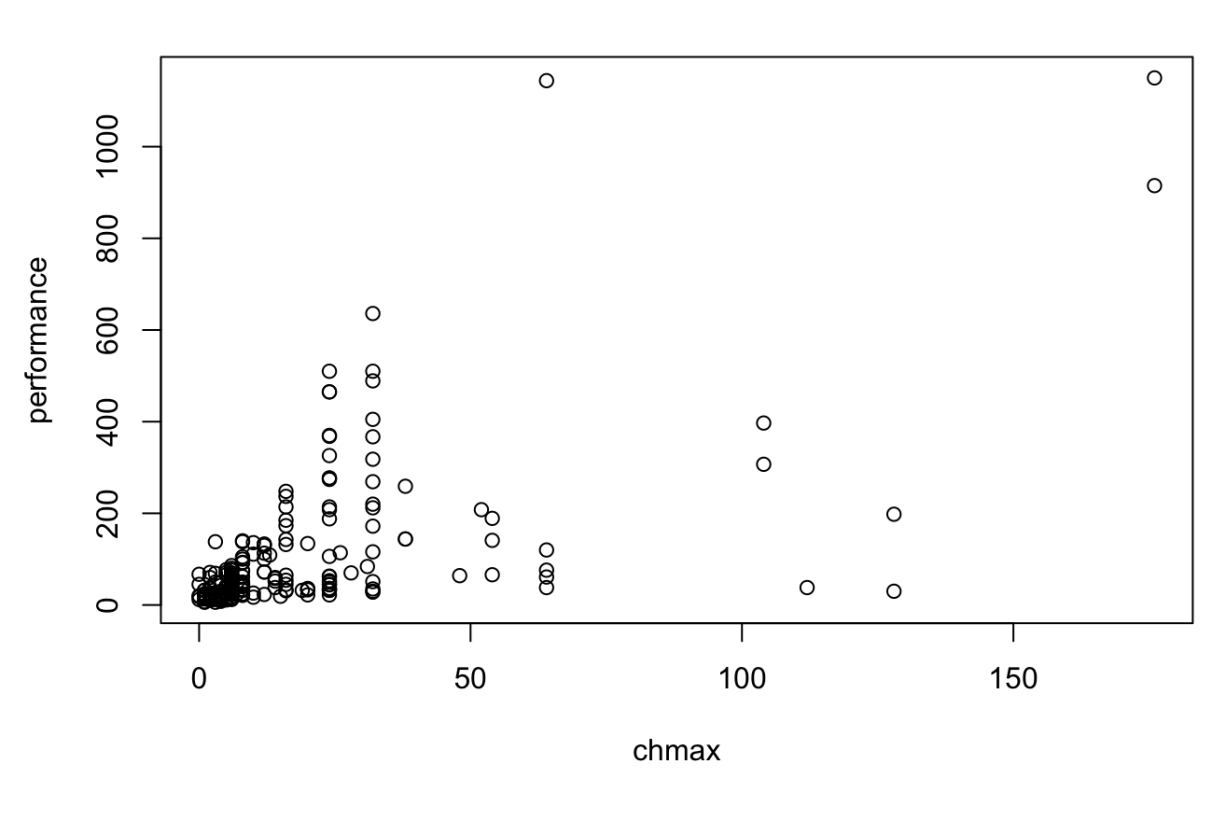
The p-value of the one sided test is so small that R interprets it as 0, so yes the β_1 is statistically significantly positive at an α-level of 0.01

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(chmax, performance)
```

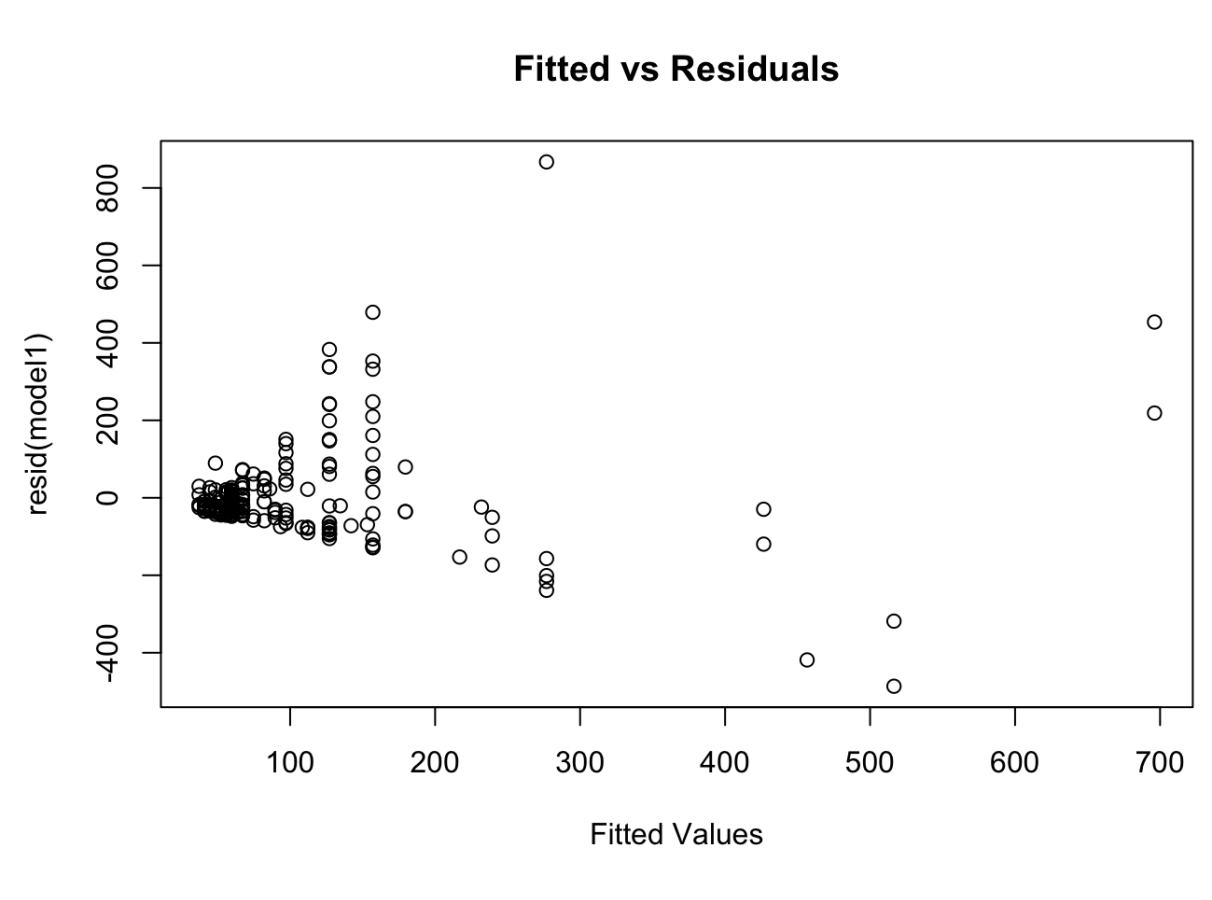


Model Assumption(s) it checks: Linearity/mean 0 assumption

Interpretation: It appears generally linear, but it is imperfect.

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")
```

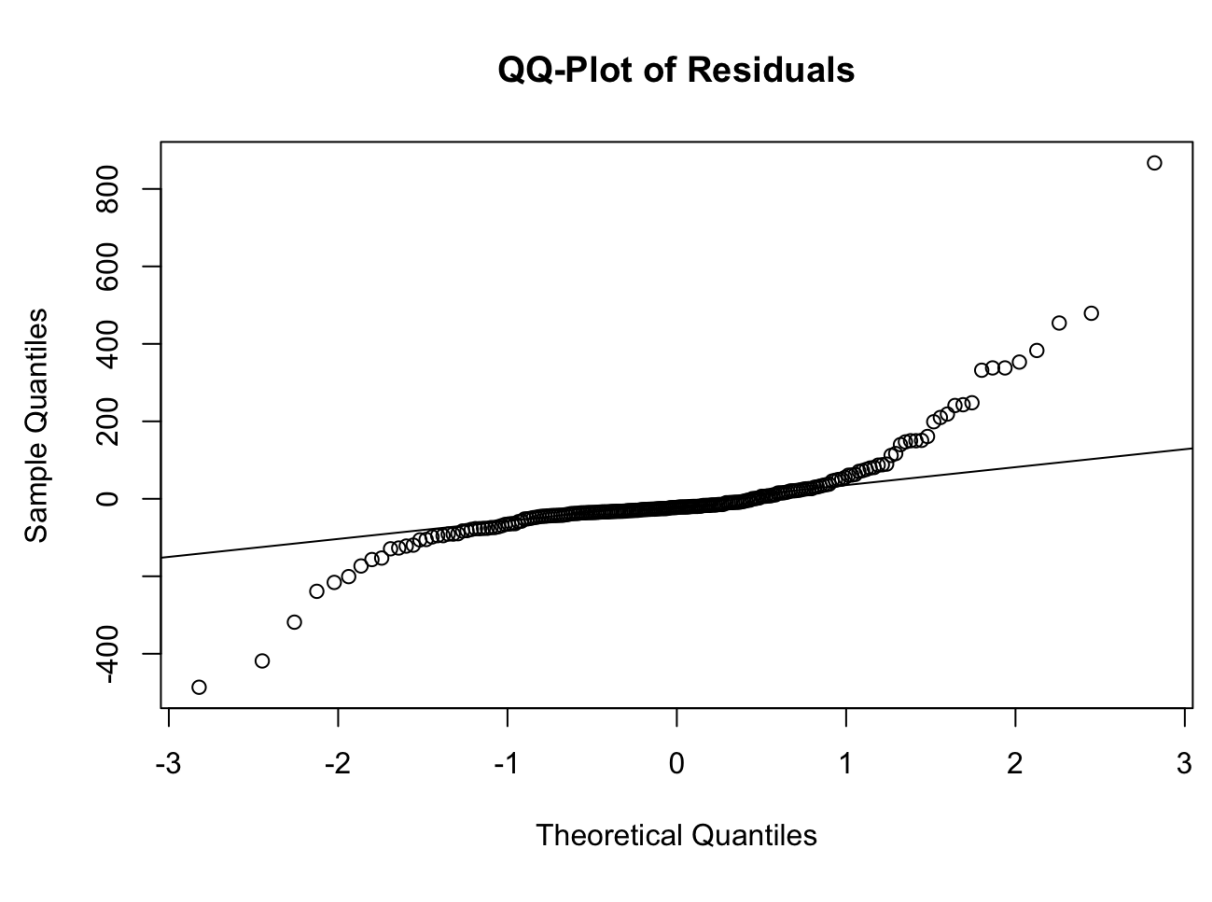


Model Assumption(s) it checks: Constant variance assumption Independence assumption (not really, we can only check for uncorrelated errors)

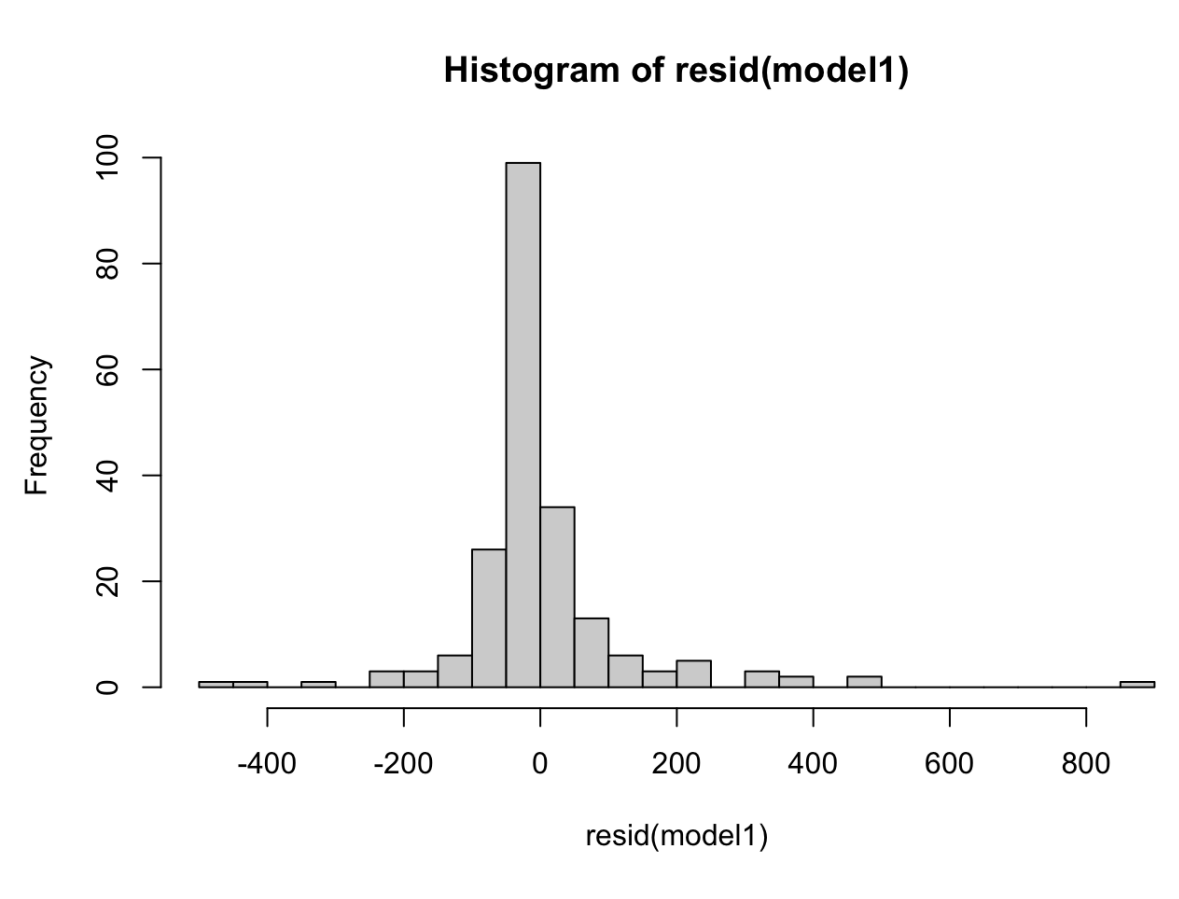
Interpretation: There is no clear shape to the distribution (such as a curve) that suggests there is a non-linear relationship, but we see a megaphone effect on the residuals, which means the constant variance assumption does not hold. There are no clusters of residuals, which suggests the errors are not correlated (related to but not proof of independence assumption)

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

```
qqnorm(resid(model1),main="QQ-Plot of Residuals")
qqline(resid(model1))
```



```
hist(resid(model1),breaks=20)
```



Model Assumption(s) it checks: Normality assumption

Interpretation: The curvature at the ends suggests non-normality. The histogram suggests a non-normal distribution that is tall at the center with long, thin tails.

Question B4: Improving the Fit - 10 pts

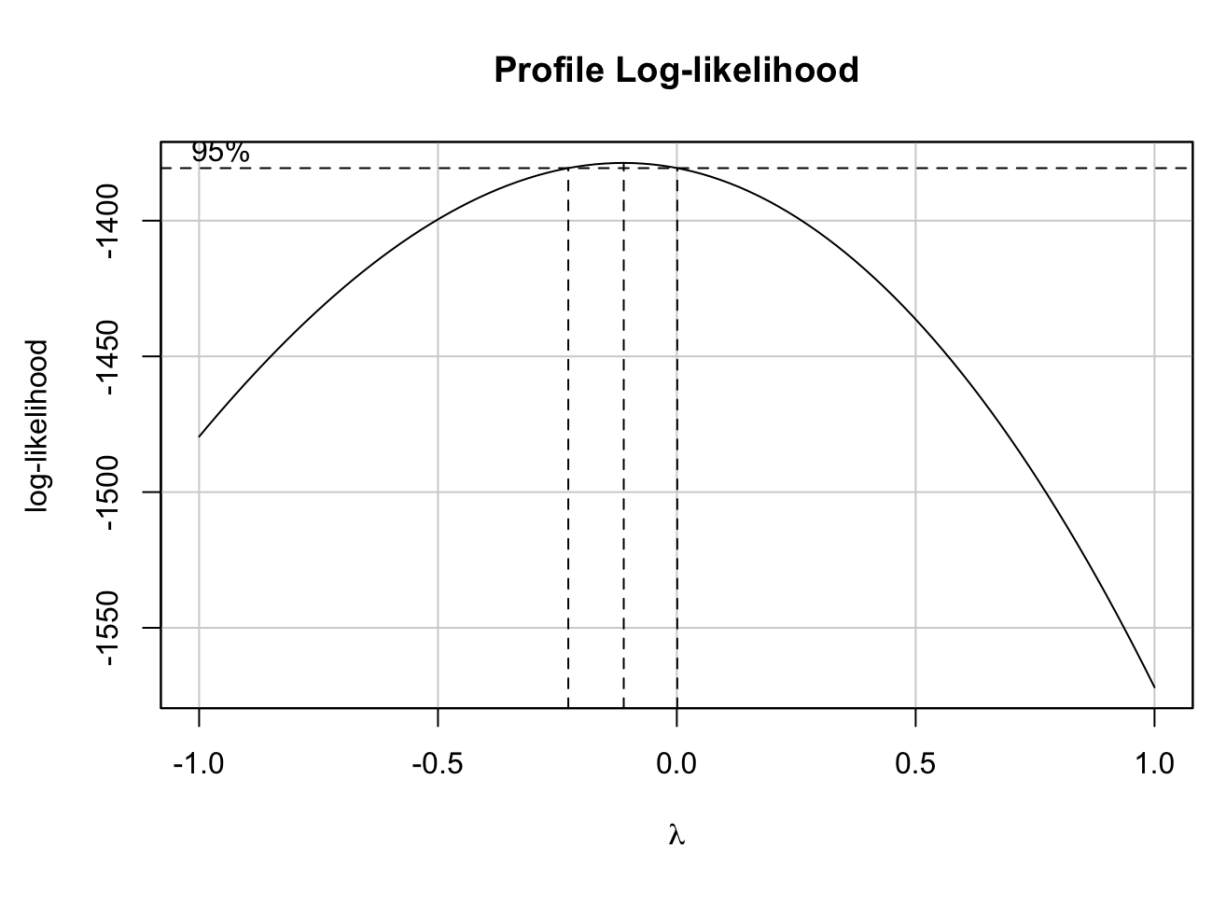
a. **2 pts** Use a Box-Cox transformation (`boxCox()`) in `car()` package or (`boxcox()`) in `MASS()` package to find the optimal λ value rounded to the nearest half integer. What transformation of the response, if any, does it suggest to perform? The optimal lambda, rounded to the nearest half integer, seems to round to 0. This suggests we use the normal logarithmic transformation according to the lecture.

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
boxCox(model1,lambda=seq(-1,1,.5))
```



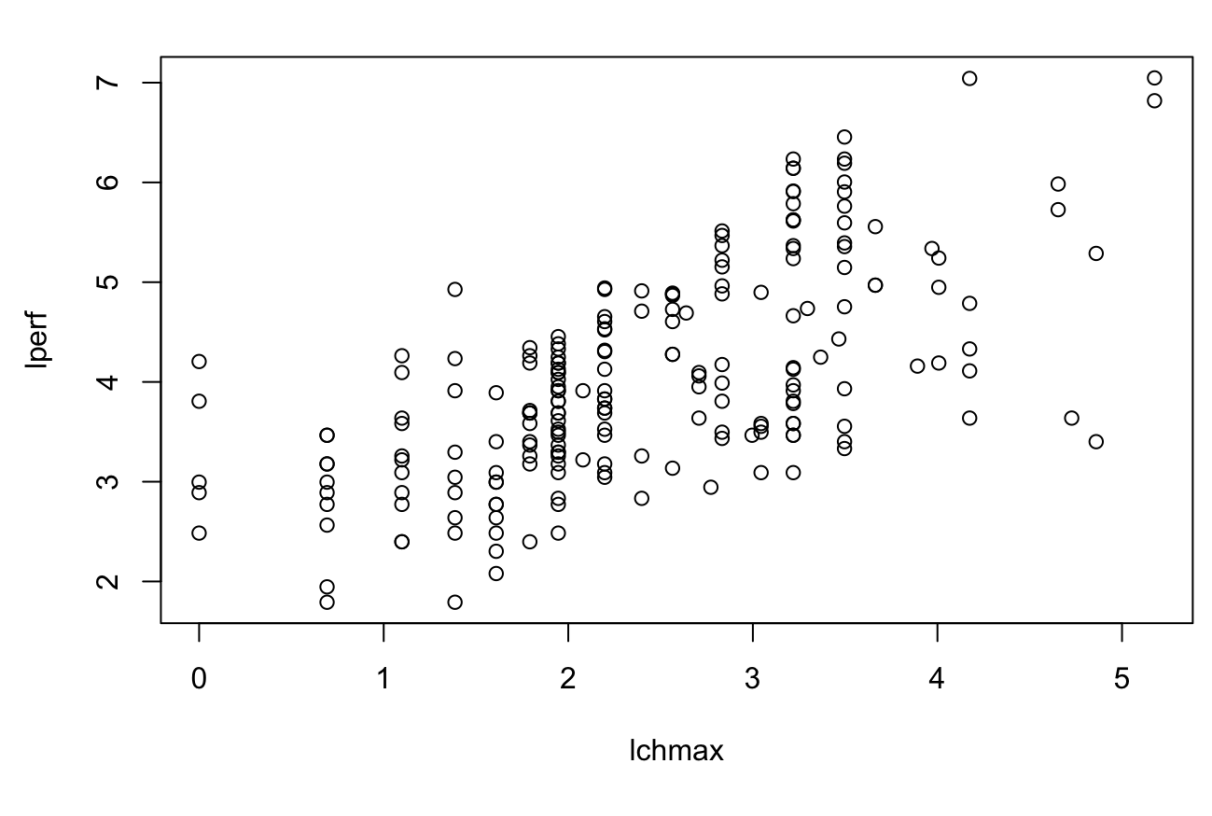
b. **2 pts** Create a linear regression model, named *model2*, that uses the log transformed *performance* as the response, and the log transformed *chmax* as the predictor. Note: The variable *chmax* has a couple of zero values which will cause problems when taking the natural log. Please add one to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

```
lchmax <- log(chmax+1)
lperf <- log(performance)
model2 = lm(lperf ~ lchmax, data)
summary(model2)
```

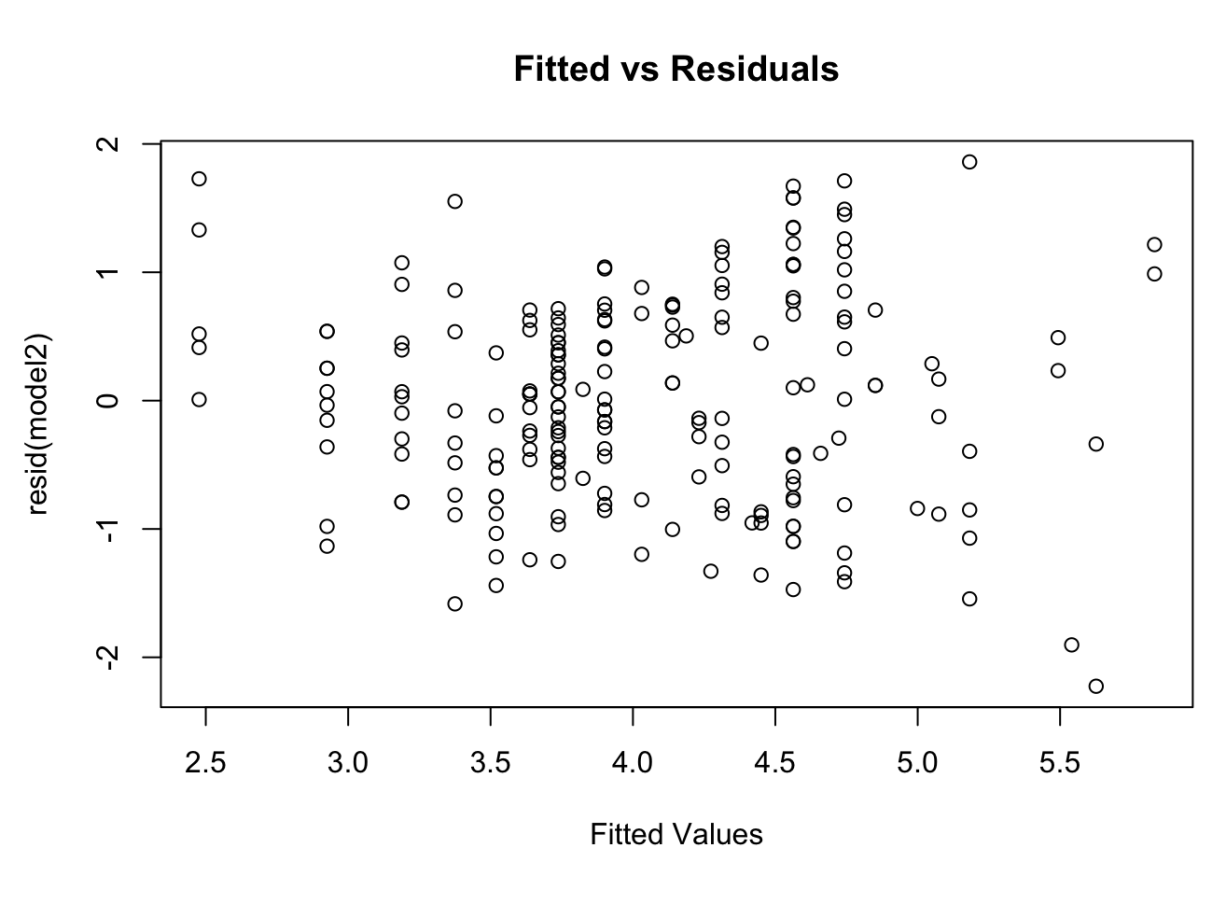
```
##
## Call:
## lm(formula = lperf ~ lchmax, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22543 -0.59429  0.01065  0.59287  1.85995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47655    0.14152    17.5   <2e-16 ***
## lchmax       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
```

c. **2 pts** Compare the R-squared values of *model1* and *model2*. Did the transformation improve the explanatory power of the model? Model 1: Multiple R-squared: 0.3663, Adjusted R-squared: 0.3632 Model 2: Multiple R-squared: 0.4103, Adjusted R-squared: 0.4074 Yes, it did improve the explanatory power somewhat
d. **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?

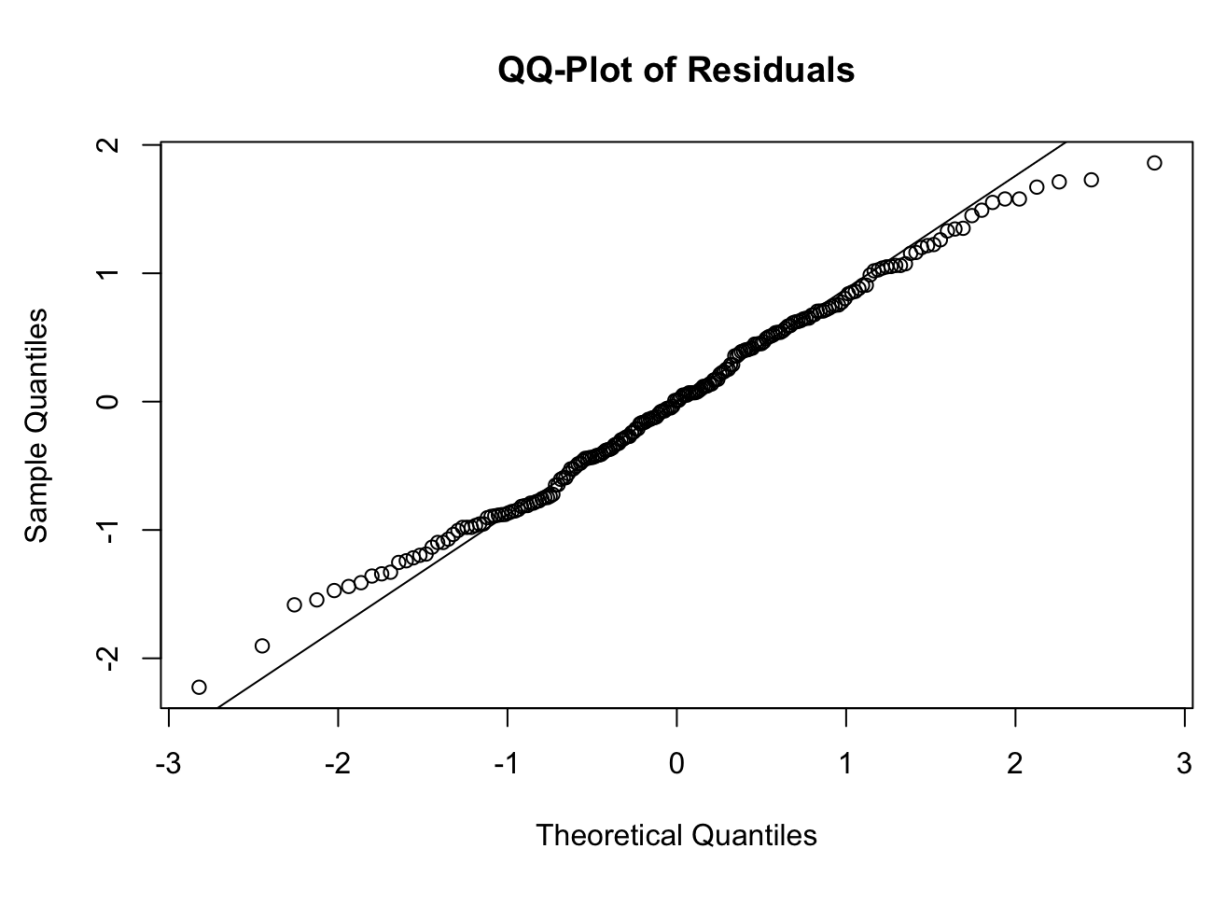
```
plot(lchmax, lperf)
```



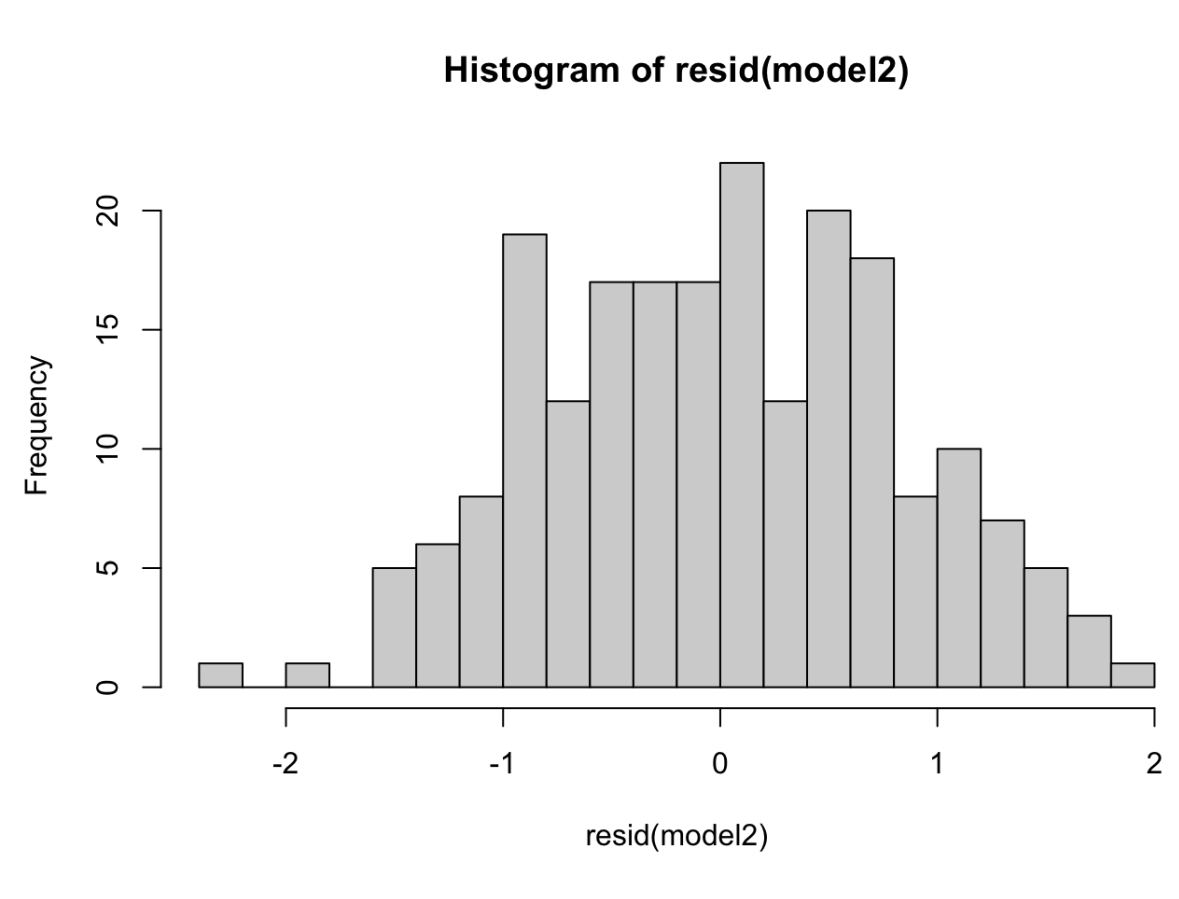
```
plot(fitted(model2),resid(model2),main="Fitted vs Residuals",xlab="Fitted Values")
```



```
qqnorm(resid(model2),main="QQ-Plot of Residuals")
qqline(resid(model2))
```



```
hist(resid(model2),breaks=20)
```



The scatterplot of log(performance) and log(chmax) shows strong linearity The scatterplot of the fitted values vs. the residual looks balanced and evenly distributed, suggesting constance variance and independence. The Q-Q plot is close to linear, with limited variance on the tails, and the histogram looks much more like a normal distribution, suggesting normality.

Based on these charts, model2 seems like a better fit than model 1

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? predict()

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

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

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

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

```
#predict(model2,log(new+1))
#buch[50]-exp(predict(model.red,new))
```

The model1 prediction is almost twice as high as the model2 prediction. Visually, when I look at the scatterplot of the original data, it seems that the leverage points that have high performance may be influencers that drag up the slope of the fitted linear regression model, resulting in larger estimates. Since model2 performed better on all the metrics we tested (better met assumptions, higher R-squared), I believe the model2 prediction to be more reliable. It is also worth noting that chmax of 128 is fairly large in the sample we trained these on, and other than 2 leverage points, all the points at this chmax or greater are much lower. Without knowing too much about the underlying problem space, it may be the case that on average, large chmax doesn't continue to lead to performance growth except in a few cases where something about the technology allows it to harness more channels, leading to those few leverage/outlier points.

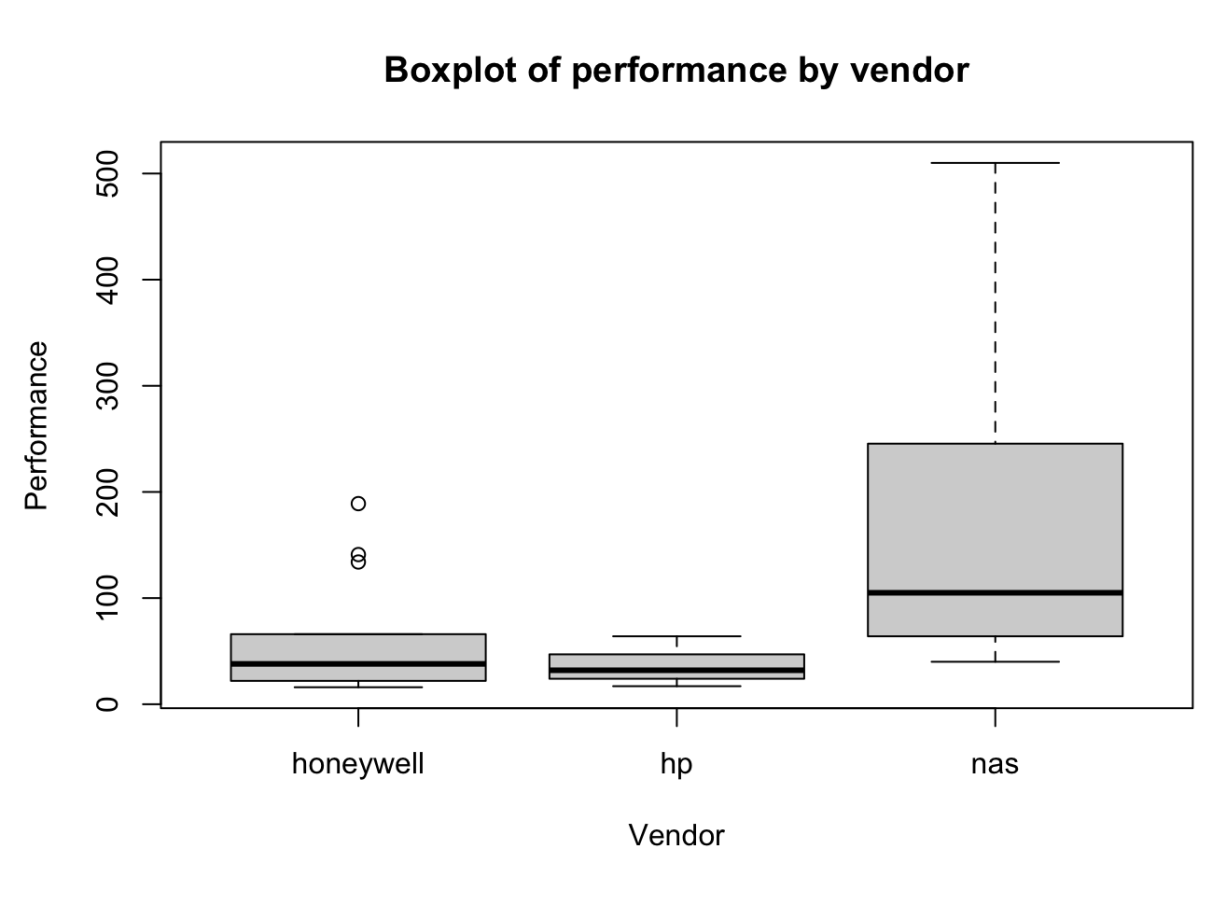
Part C. ANOVA - 8 pts

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$performance~as.factor(data2$vendor), main="Boxplot of performance by vendor",xlab="Vendor",ylab="Pe
rformance")
```



The boxplots suggest that hp has the lowest performance products, honeywell is somewhat higher with a few outliers with moderate performance, and nas specializes in high performance models, including the extreme outliers.

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.

```
model3 = aov(performance ~ vendor,data = data2)
summary(model3)
```

```
##             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(model3,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
```

Yes, the P value of the F-test is .00553, which is significant at an alpha level of .05. This means that there is strong statistical evidence that at least one of the means is different than the others.

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(model3, conf.level=.95)
```

```
##   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
```

nas-honeywell and nas-hp are statistically significantly different from each other according to the Tukey pairwise comparison.