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	7.224	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.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.

Answer A1

model <- aov(Phase ~ Treatment, data = data)
summary(model)
See above for completed 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 (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.

Answer A2

μ_1, μ_2, and μ_3 is the population mean value for each group: Control, Knees, and Eyes respectively.

control <- c(0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27)
knees <- c(0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61)
eyes <- c(-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83)

u1 <- mean(control) = -0.30875
u2 <- mean(knees) = -0.3357143
u3 <- mean(eyes) = -1.551429

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_o: $\mu_1 = \mu_2 = \mu_3$

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

H_1: The population means are not all equal

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

$F(2, 19)$

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

Pr(>F) = 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.

Yes, it does affect phase shift. Since the p-value is less than alpha, you reject the null hypothesis.

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()`.

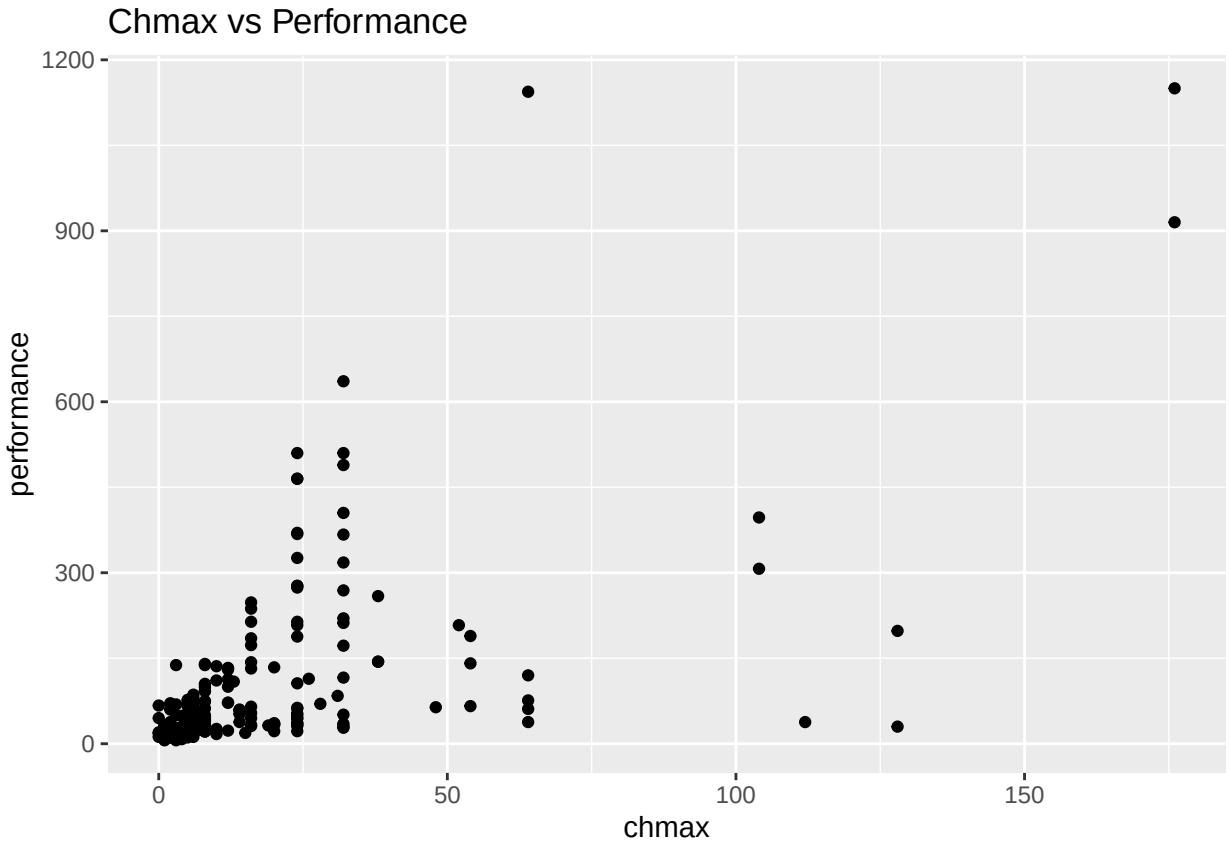
```
# Read in the data
data = read.csv("machine.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(data, 3)
```

```
##    vendor chmax performance
## 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.

```
library(ggplot2)
machine <- read.csv("~/Desktop/gatech/Fall 2022/ISYE6414(Regression)/Module 1/Homework 1/machine.csv")
plot <- ggplot(machine, aes(x=chmax, y=performance)) + geom_point()
plot + ggtitle('Chmax vs Performance')
```



The general trend of the scatter plot shows a positive relationship. As the number of channels increases we see an increase in performance as well.

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$performance, machine$chmax)
```

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

Since is correlation coefficient value is 0.6, it can be interpreted as having a strong positive relationship.

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

Yes, I would recommend a linear regression model as it has a positive correlation coefficient.

d. **1 pts** Based on the analysis above, would you pursue a transformation of the data? *Do not transform the data.*

No, not yet since we need to plot the residuals to see if the points are non-linear.

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?

$\beta_0 = 37.2252$ and $\beta_1 = 3.7441$

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

$y = 37.2252 + 3.7441x$

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

For every one unit of channels added to the CPU will increase its performance by 3.7441 units.

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

```
confint(model1)
```

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

β_1 has a confidence interval of [3.069251, 4.418926]. alpha = 1 - confidence level = 0.05. β_1 has a p-value less than the alpha value it is statistically significant.

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?

Even if the $\alpha = 0.01$, the p-value of 2e-16 is lower, therefore, β_1 is significant.

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

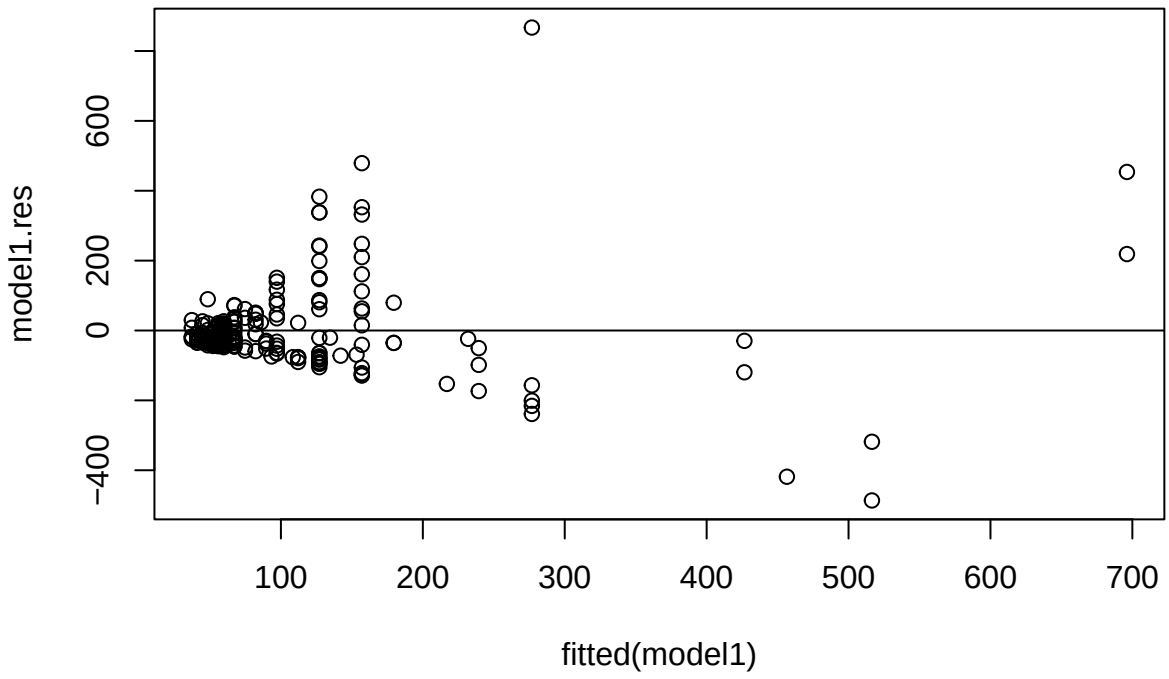
```
library(ggplot2)
model1 = lm(performance ~ chmax, data)
model1.res = resid(model1)
plot <- ggplot(model1, aes(x=chmax, y=performance)) + geom_point()
```

Model Assumption(s) it checks: Checks for Linearity

Interpretation: The plot depicts some positive correlation and little linearity.

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

```
plot(fitted(model1), model1.res)
abline(0,0)
```

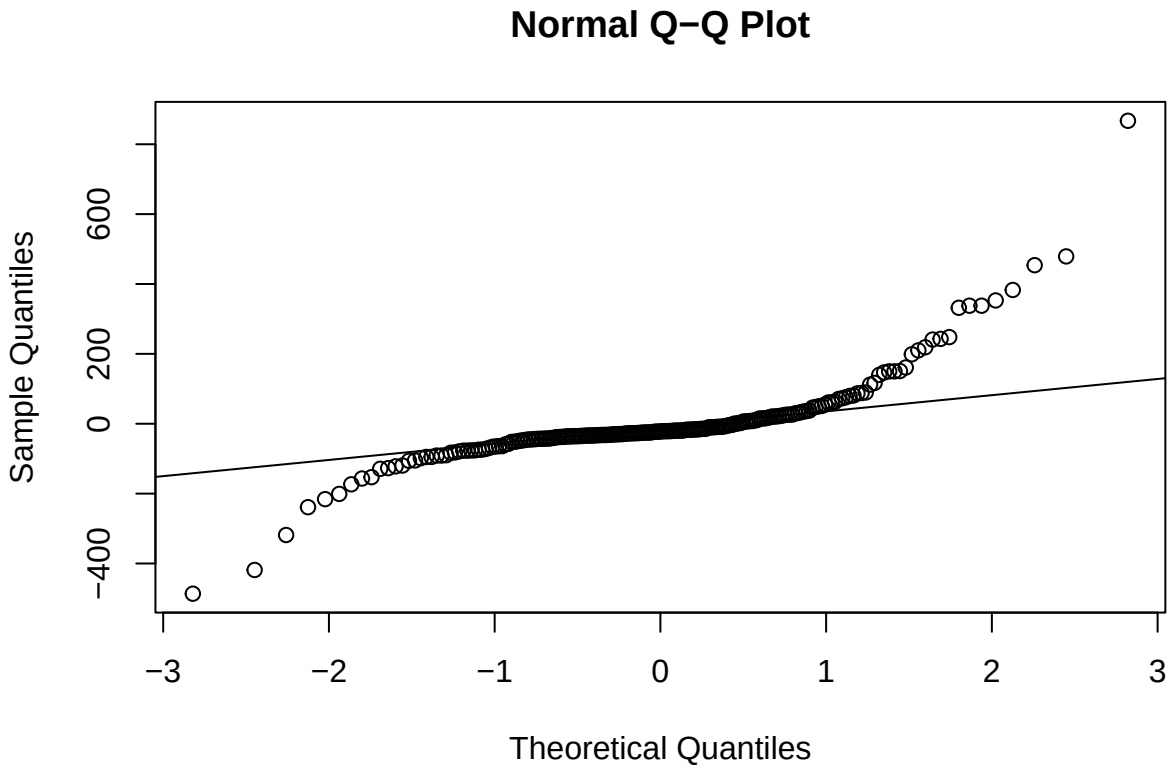


Model Assumption(s) it checks: Checks for Independence and Homoscedasticity

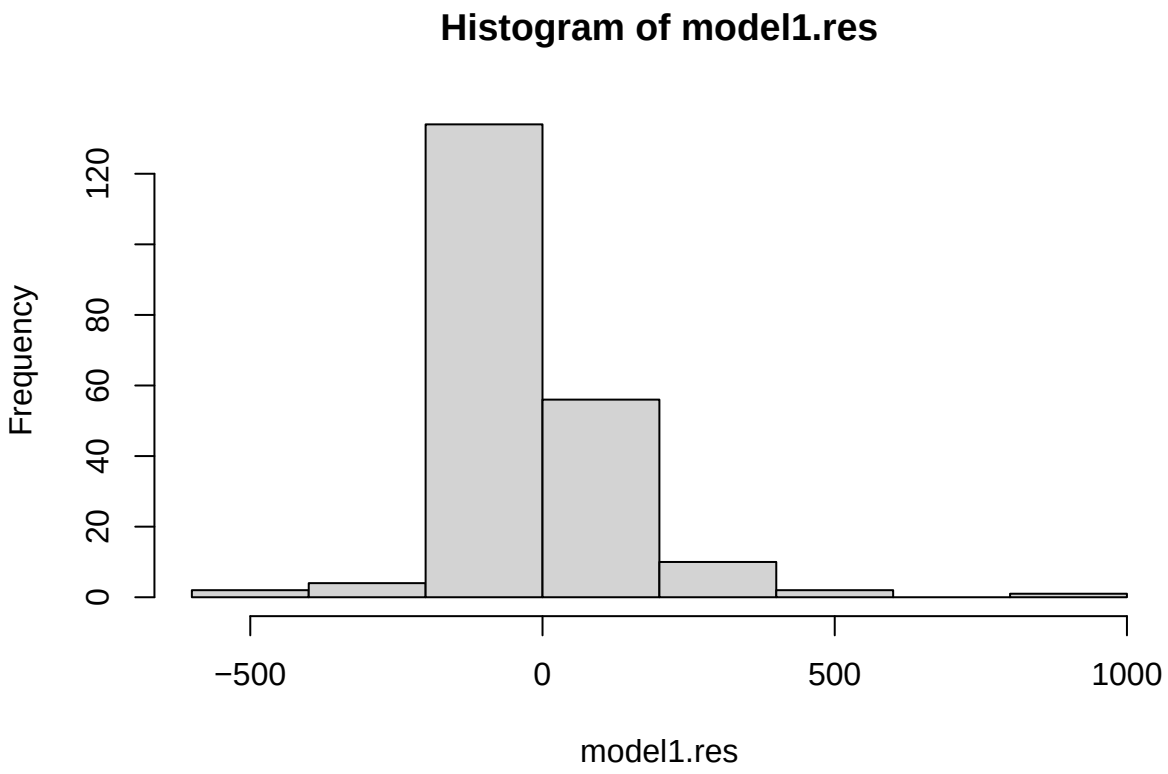
Interpretation: The plot depicts that the data is not homoscedastic and not independent.

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

```
# Your code here...
qqnorm(model1.res)
qqline(model1.res)
```




```
hist(model1.res)
```



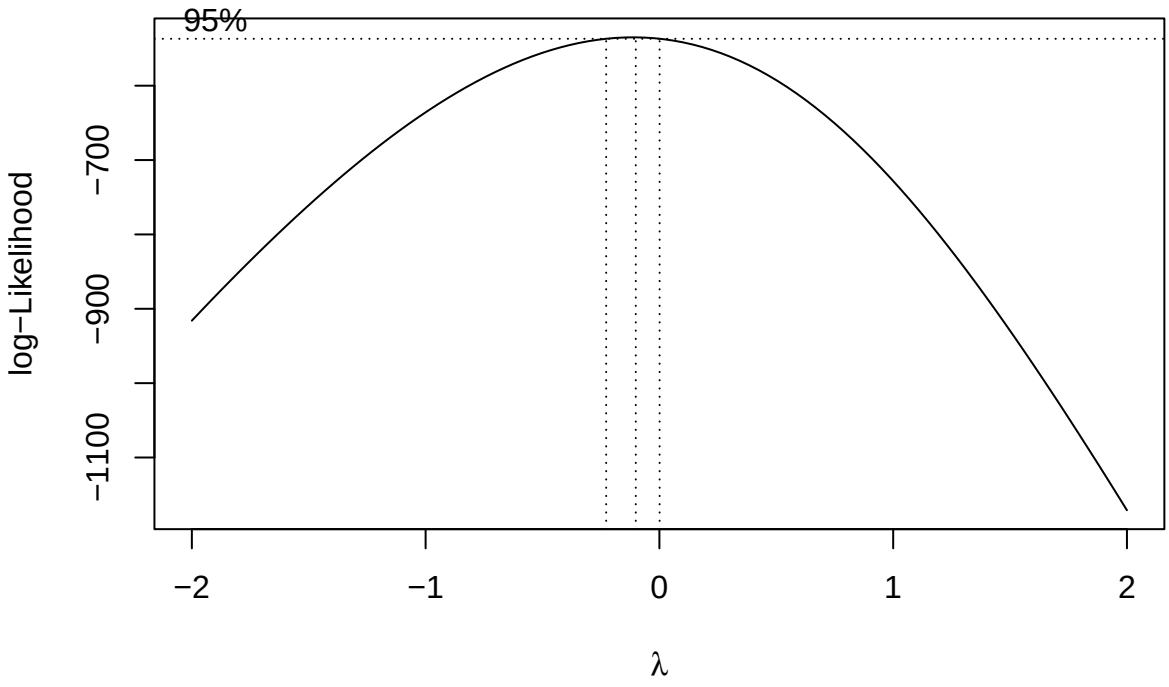
Model Assumption(s) it checks: Checks for Normality

Interpretation: The plot depicts that the residuals are not normally distributed, so the data needs to be transformed.

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?

```
library(MASS)
x = machine[,2]
y = machine[,3]
model1.bc <- boxcox(y ~ x)
```




```
lambda <- model1.bc$x[which.max(model1.bc$y)]
signif(lambda, digits = 2)
```

```
## [1] -0.1
```

After using a boxcox transmation on the data, we are given an optimal λ = -0.1, which can be rounded to zero. If λ is 0, then we will need to perform a 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

```
# Your code here...
chmax_log <- log(machine[,2] + 1)
perf_log <- log(machine[,3])
model2 <- lm(perf_log ~ chmax_log)
summary(model2)
```

```
##
## Call:
## lm(formula = perf_log ~ chmax_log)
##
## 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_log    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
```

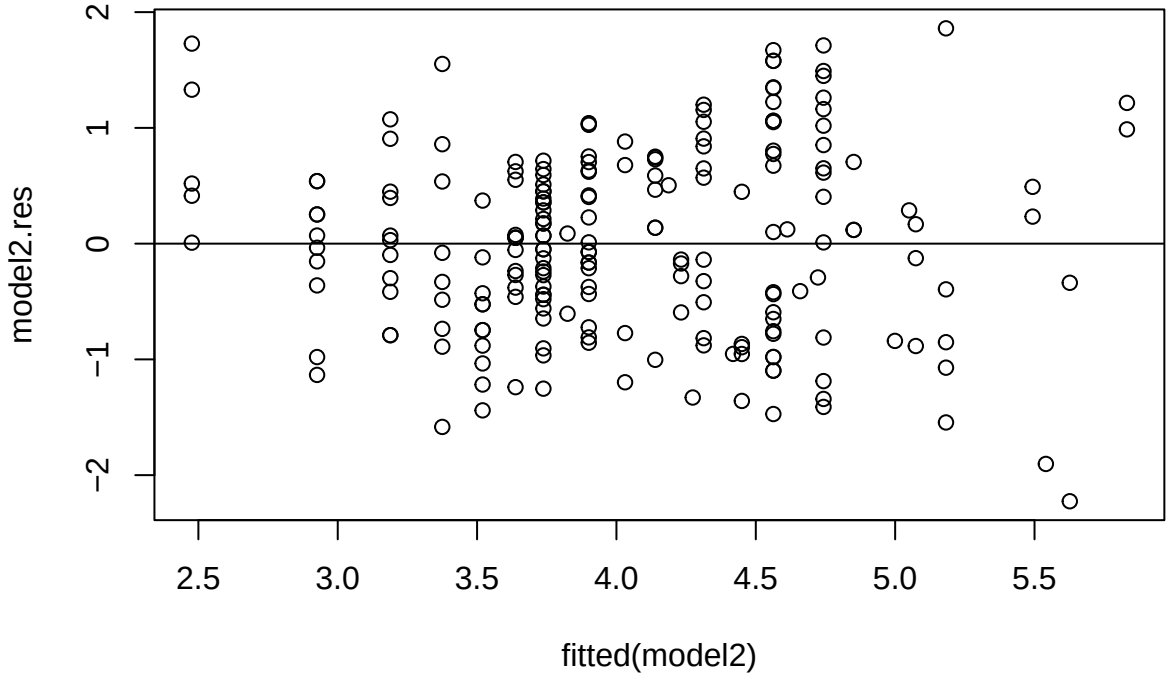
c. **2 pts** Compare the R-squared values of *model1* and *model2*. Did the transformation improve the explanatory power of the model?

model1 $R^2 = 0.3663$ and model2 $R^2 = 0.4103$. Since the R^2 value increased the explanatory power has improved.

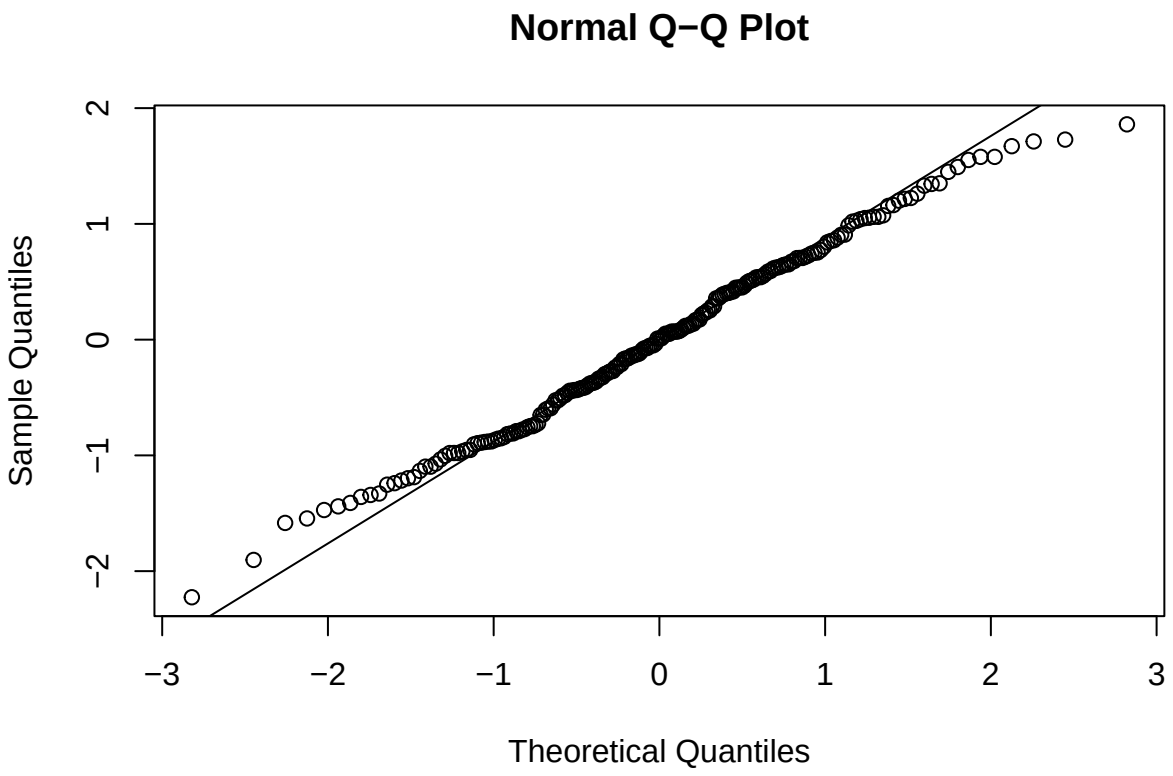
d. **4 pts** Similar to Question B3, assess and interpret all model assumptions of *model2*. A model is considered a good fit if all assumptions hold. Based on your interpretation of the model assumptions, is *model2* a good fit?

```
# Your code here..
library(ggplot2)
model2.res = resid(model2)
plot2 <- ggplot(model2, aes(x=chmax_log, y=perf_log)) + geom_point()

plot(fitted(model2), model2.res)
abline(0,0)
```




```
qqnorm(model2.res)
qqline(model2.res)
```




```
hist(model2.res)
```

Yes, model2 is a good fit.

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...
predict(model1, data.frame(chmax = 128), interval = "prediction", level = 0.95)
```

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

```
predict(model2, data.frame(chmax_log = log(128)), interval = "prediction", level = 0.95)
```

```
##       fit      lwr      upr
## 1 5.62158 4.005673 7.237486
```

The performance prediction value for model 1 does not fit well as the value is greater than the given data and plot. Whereas the prediction value for model 2 fits better given the log(data) and plot.

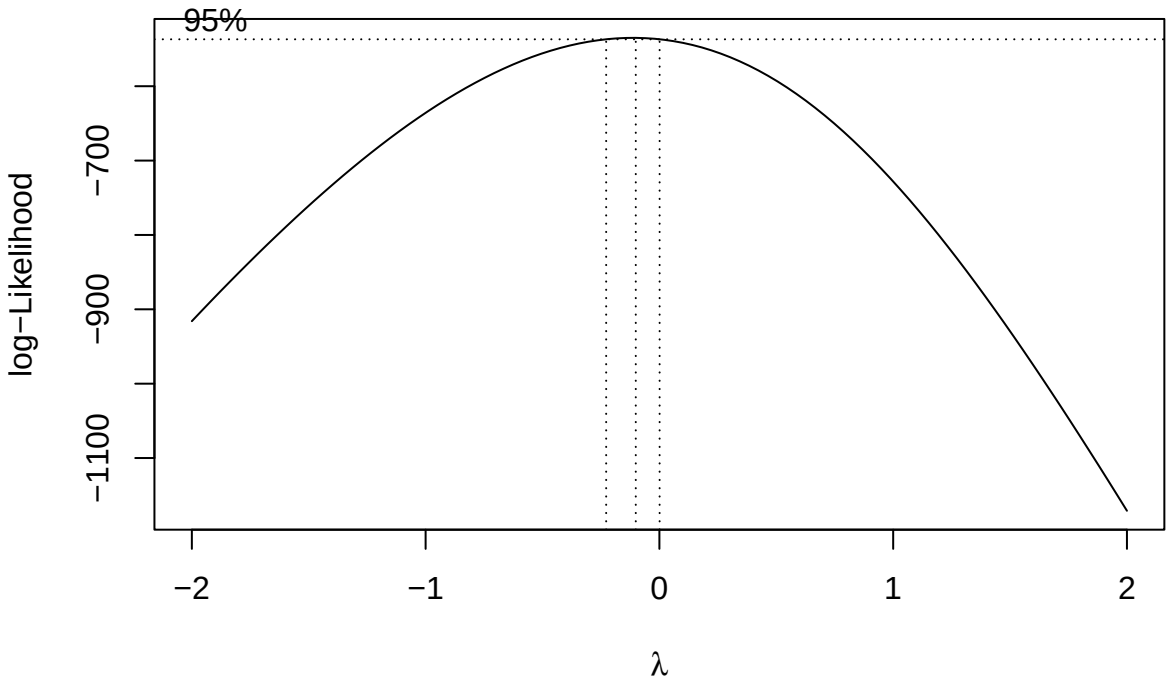
Part C. ANOVA - 8 pts

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
machine2 = machine[machine$vendor %in% c("honeywell", "hp", "nas"), ]
machine$vendor = factor(machine$vendor)
```

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

```
x2 <- machine2[,1]
y2 <- machine2[,3]
machine2.bc <- boxcox(y ~ x)
```




```
(lambda2 <- machine2.bc$x[which.max(machine2.bc$y)])
```

```
## [1] -0.1010101
```

The lambda value rounds to zero, so a log transformation is needed.

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.

```r
machine.aov <- aov(performance ~ vendor, data = machine2)
summary(machine.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
```

Yes, you can reject the null hypothesis since the p-value $= 0.00553 < \alpha = 0.05$.

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?

```r
# Your code here...
TukeyHSD(machine.aov, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = performance ~ vendor, data = machine2)
##
## $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
```

Based on the Tukey pairwise comparison, nas-honeywell and nas-hp are statistically significantly different as the difference value is the furthest from 0 and the p-adj value is less than the alpha value 0.05, so you reject the null hypothesis.