# hw7

Xinrui Hu

2023-04-04

Question 1.1 Validate that the data contains 28 observations. Include only observations with no missing values

```
animalsDf <- read.csv("AnimalsStat.csv")
nrow(animalsDf [complete.cases(animalsDf ), ])
```

```
## [1] 28
```

```
nrow (animalsDf)
```

```
## [1] 28
```

Question 1.2 Show the first observations of the table

```
head(animalsDf)
```

```
##              Name     Body Brain
## 1 Mountain beaver     1.35   8.1
## 2             Cow   465.00 423.0
## 3       Grey wolf    36.33 119.5
## 4            Goat    27.66 115.0
## 5      Guinea pig     1.04   5.5
## 6     Dipliodocus 11700.00  50.0
```

Question 1.3 Use the summary function to answer the following questions: What is the average body mass? What is the average brain mass? Does the data contain any animals with a body mass of 0 kg Does the data contain any animals with a brain mass of 0 g Does the data contain any missing values?

```
summary(animalsDf)
```

```
##      Name                Body              Brain
##  Length:28          Min.   :    0.02   Min.   :   0.40
##  Class :character   1st Qu.:    3.10   1st Qu.:  22.23
##  Mode  :character   Median :   53.83   Median : 137.00
##                     Mean   : 4278.44   Mean   : 574.52
##                     3rd Qu.:  479.00   3rd Qu.: 420.00
##                     Max.   :87000.00   Max.   :5712.00
```

```
# What is the average body mass? 4278.44kg
# What is the average brain mass? 574.52g
# Does the data contain any animals with a body mass of 0 kg ? No
# Does the data contain any animals with a brain mass of 0 g ? No
# Does the data contain any missing values? No
```

Question 1.4 Which animal has the smallest and which has the largest brain mass in the dataset?

```
animalsDf[which.min(animalsDf$Brain),] # Mouse
```

```
##     Name  Body Brain
## 20 Mouse 0.023   0.4
```

```
animalsDf[which.max(animalsDf$Brain),] # African elephant
```

```
##                Name Body Brain
## 15 African elephant 6654  5712
```

Question 1.5 Create a new variable for the brain-to-body mass ratio (i.e., brain mass / body mass)

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
animalsDf <- animalsDf %>%  mutate(ratio = animalsDf$Brain / animalsDf$Body)
```

Question 1.6 Which animal has the smallest ratio, and which has the largest?

```
animalsDf[which.min(animalsDf$ratio),] # Brachiosaurus
```

```
##             Name  Body Brain       ratio
## 26 Brachiosaurus 87000 154.5 0.001775862
```

```
animalsDf[which.max(animalsDf$ratio),] # Rhesus monkey
```

```
##             Name Body Brain    ratio
## 17 Rhesus monkey  6.8   179 26.32353
```

Question 1.7 Create the following four regressions: 1. Body mass as the predictor and brain mass as the outcome 2. Log body mass as the predictor and brain mass as the outcome 3. Body mass as the predictor and log brain mass as the outcome 4. Log body mass as the predictor and log brain mass as the outcome Show the summary for each model. Which one has the largest R2?

```
model1 <- lm(Brain ~ Body, data = animalsDf)
model2 <- lm(Brain ~ log(Body), data = animalsDf)
model3 <- lm(log(Brain) ~ Body, data = animalsDf)
model4 <- lm(log(Brain) ~ log(Body), data = animalsDf)
summary(model1)
```

```
##
## Call:
## lm(formula = Brain ~ Body, data = animalsDf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -576.0 -554.1 -438.1 -156.3 5138.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.764e+02  2.659e+02   2.168   0.0395 *
```

```
## Body        -4.326e-04  1.589e-02  -0.027   0.9785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1360 on 26 degrees of freedom
## Multiple R-squared:  2.853e-05,  Adjusted R-squared:  -0.03843
## F-statistic: 0.0007417 on 1 and 26 DF,  p-value: 0.9785
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = Brain ~ log(Body), data = animalsDf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1511.4  -447.9  -251.0    12.9  4415.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.13     335.43   0.099   0.9221
## log(Body)     143.55      63.47   2.262   0.0323 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1244 on 26 degrees of freedom
## Multiple R-squared:  0.1644, Adjusted R-squared:  0.1323
## F-statistic: 5.116 on 1 and 26 DF,  p-value: 0.0323
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = log(Brain) ~ Body, data = animalsDf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2903 -1.3188  0.3899  1.6632  4.1963
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.374e+00  4.763e-01   9.183 1.21e-09 ***
## Body        1.203e-05  2.845e-05   0.423    0.676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.437 on 26 degrees of freedom
## Multiple R-squared:  0.006827,   Adjusted R-squared:  -0.03137
## F-statistic: 0.1787 on 1 and 26 DF,  p-value: 0.676
```

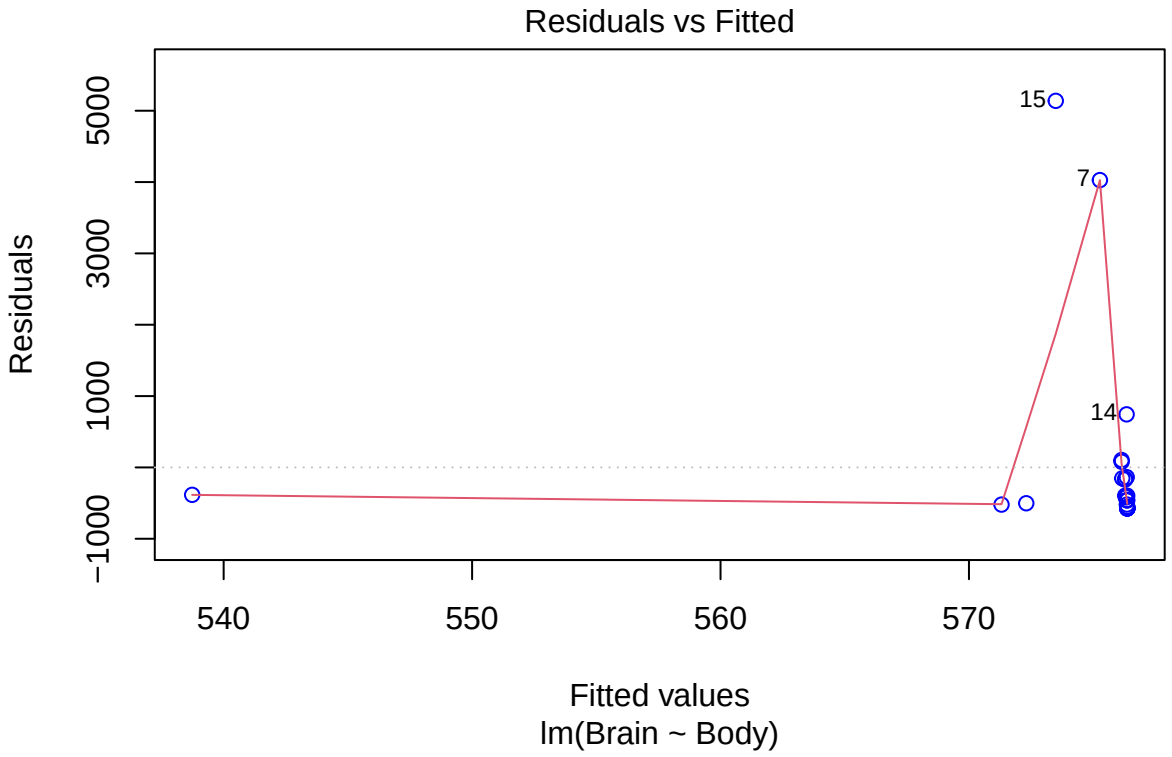
```
# The model4 has the greatest R2
summary(model4)
```

```
##
## Call:
## lm(formula = log(Brain) ~ log(Body), data = animalsDf)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2890 -0.6763  0.3316  0.8646  2.5835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55490    0.41314   6.184 1.53e-06 ***
## log(Body)    0.49599    0.07817   6.345 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 26 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.5925
## F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

Question 1.8 Create a scatter plot with a regression line and a residuals vs fitted plot for the best regression.

```
plot(model1, which=1, col = "blue")
```

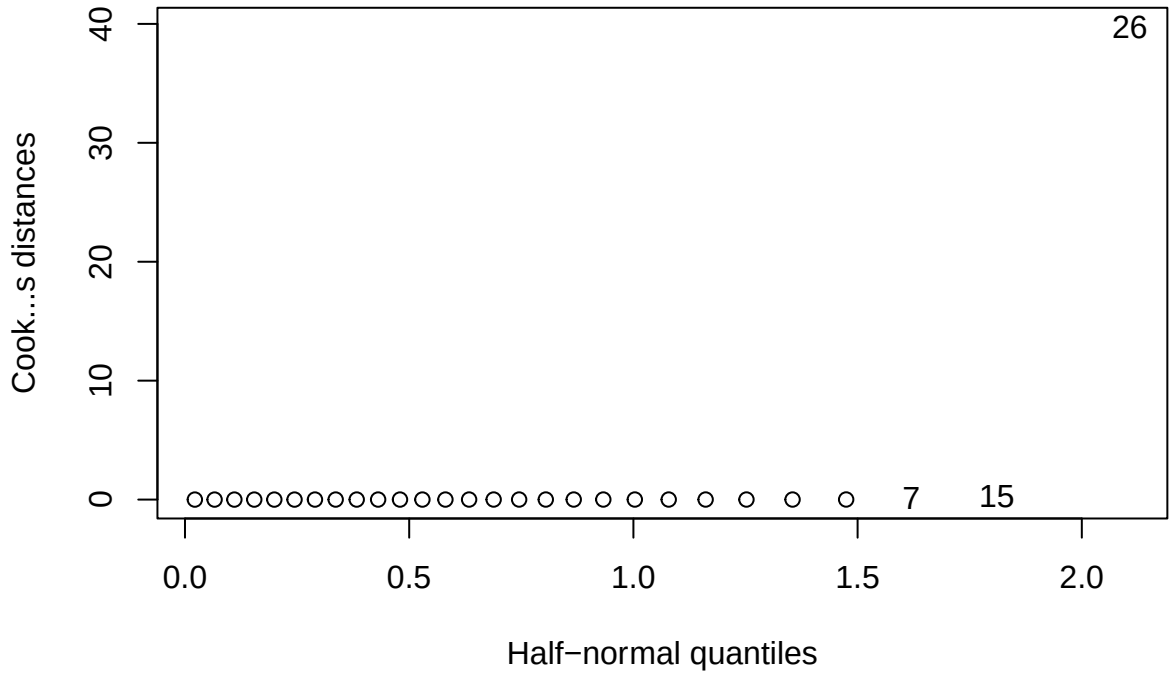



Question 1.9 Based on the best model you found, Identify influential points using a half-normal plot of the Cook's distances. Which three animals have the largest distances?

```
library(faraway)
cook <- cooks.distance(model1)
cook[which(cook > 0.5)]
```

```
##       26
## 39.76602
```

```
Animals <- row.names(animalsDf)
halfnorm(cook,3,labs = Animals , ylab = "Cook's distances")
```



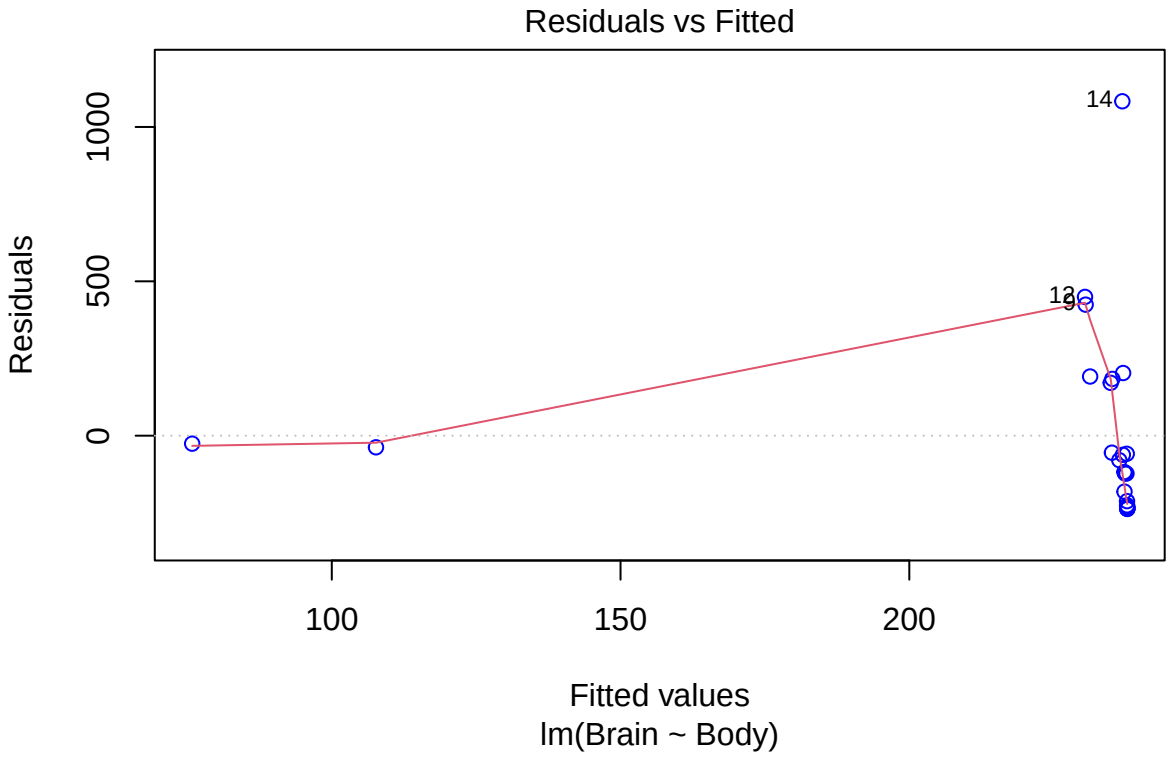

```
# Asian elephant, African elephant, Brachiosaurus have the largest distance
```

Question 1.10 Create a new data frame with the three observations removed

```
df <- animalsDf[-c(7, 15, 26),]
```

Question 1.11 Fit a new model based on the new data frame (with the three points removed). Show again the type of plots you created for question 1.8

```
lmod<- lm(Brain ~ Body, data = df)
plot(lmod, which=1, col = "blue")
```




Question 2.1 Fit a model with O3 as the outcome and temp, humidity, ibh as predictors. Which coefficients are significant at the 5% level?

```
mod <- lm(O3 ~ temp + humidity + ibh, data = ozone)
summary(mod)
```

```
##
## Call:
## lm(formula = O3 ~ temp + humidity + ibh, data = ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5291  -3.0137  -0.2249   2.8239  13.9303
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.049e+01  1.616e+00  -6.492 3.16e-10 ***
## temp         3.296e-01  2.109e-02  15.626  < 2e-16 ***
## humidity     7.738e-02  1.339e-02   5.777 1.77e-08 ***
## ibh         -1.004e-03  1.639e-04  -6.130 2.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.524 on 326 degrees of freedom
## Multiple R-squared:  0.684,  Adjusted R-squared:  0.6811
## F-statistic: 235.2 on 3 and 326 DF,  p-value: < 2.2e-16
```

```
# all the predictors are significant at 5% level
```

Question 2.2 Fit the same model but add an interaction between temp and humidity. Has the adjusted R2 increased compared to the previous model? Is the interaction coefficient significant? Is the temp variable

significant? Should we remove the temp variable while keeping the interaction?

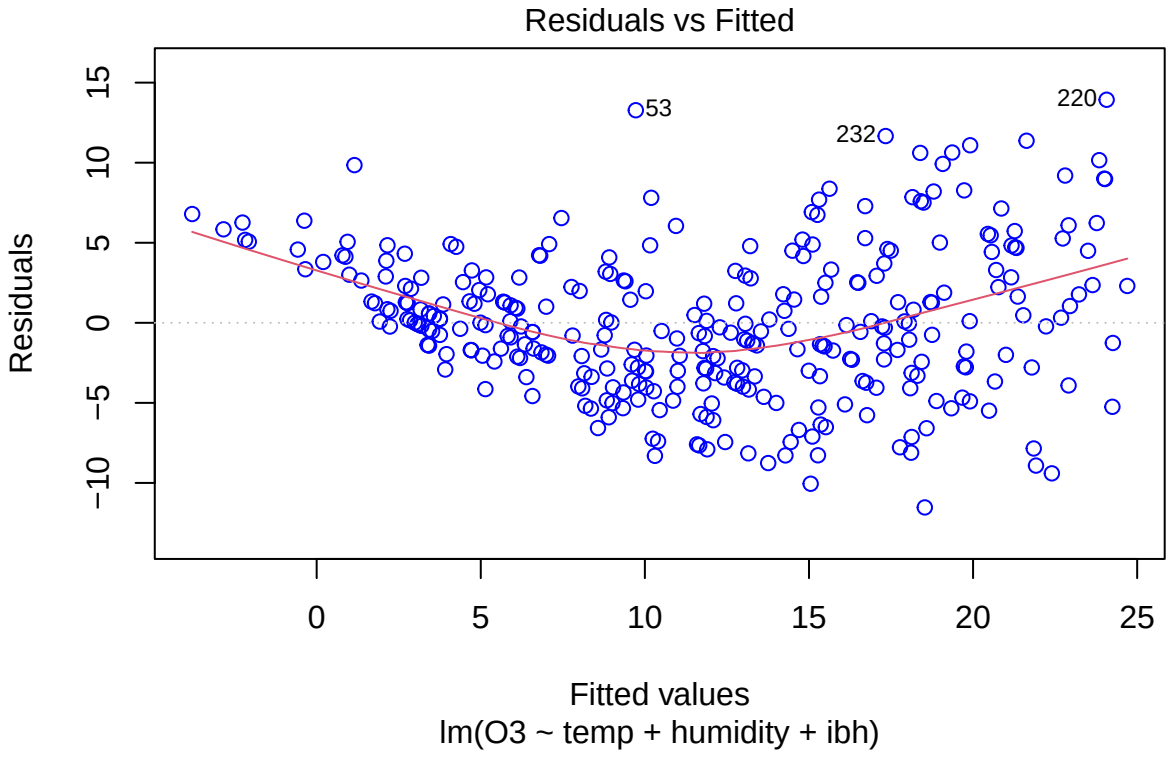
```
mod1 <- lm(O3 ~ temp + humidity + ibh + temp * humidity, data = ozone)
summary(mod1)
```

```
##
## Call:
## lm(formula = O3 ~ temp + humidity + ibh + temp * humidity, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.204  -2.890  -0.176   2.508  14.476
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.9318952  4.0129533   2.724   0.0068 **
## temp          -0.0479114  0.0683146  -0.701   0.4836
## humidity      -0.2741679  0.0621176  -4.414 1.38e-05 ***
## ibh           -0.0010115  0.0001563  -6.472 3.56e-10 ***
## temp:humidity  0.0060593  0.0010478   5.783 1.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.315 on 325 degrees of freedom
## Multiple R-squared:  0.7135, Adjusted R-squared:  0.7099
## F-statistic: 202.3 on 4 and 325 DF,  p-value: < 2.2e-16
```

```
# R2 increased compare with the prior model, interaction model is significant, temp is not significant,
```

Question 2.3 Create a residuals vs fitted plot. Do you detect any issues?

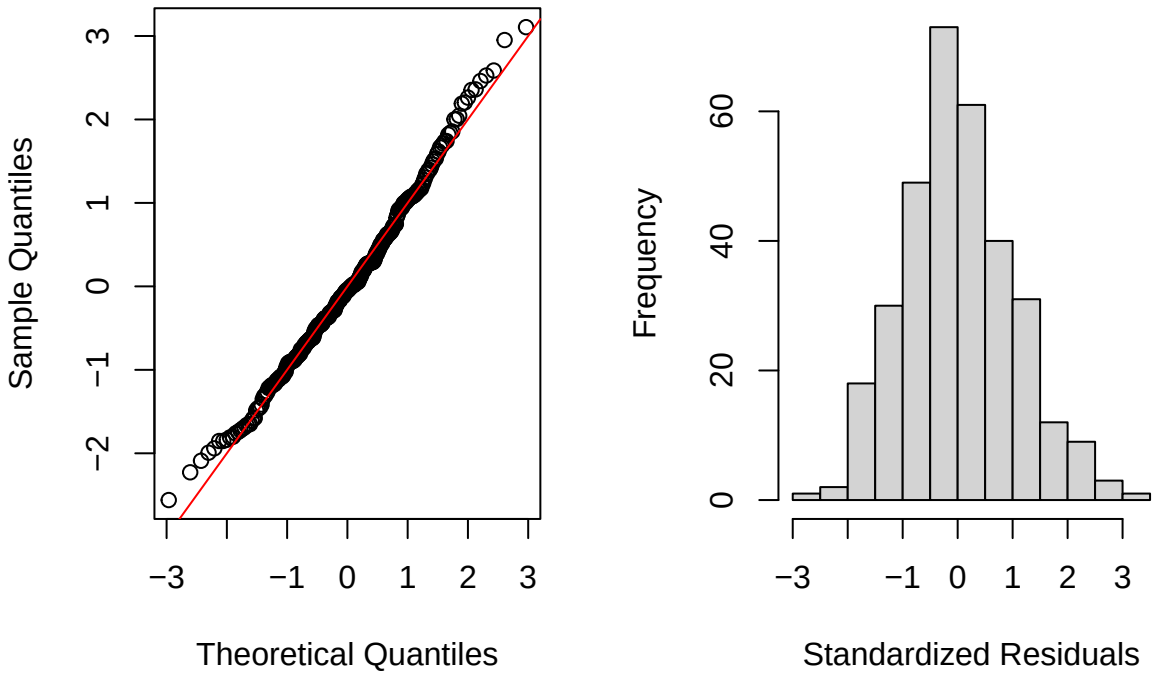
```
plot(mod, which = 1, col = "blue")
```



```
# The plot is not linearly
```

Question 2.4 Create a Q-Q plot and histogram of the standardized residuals. Do you see any issues?

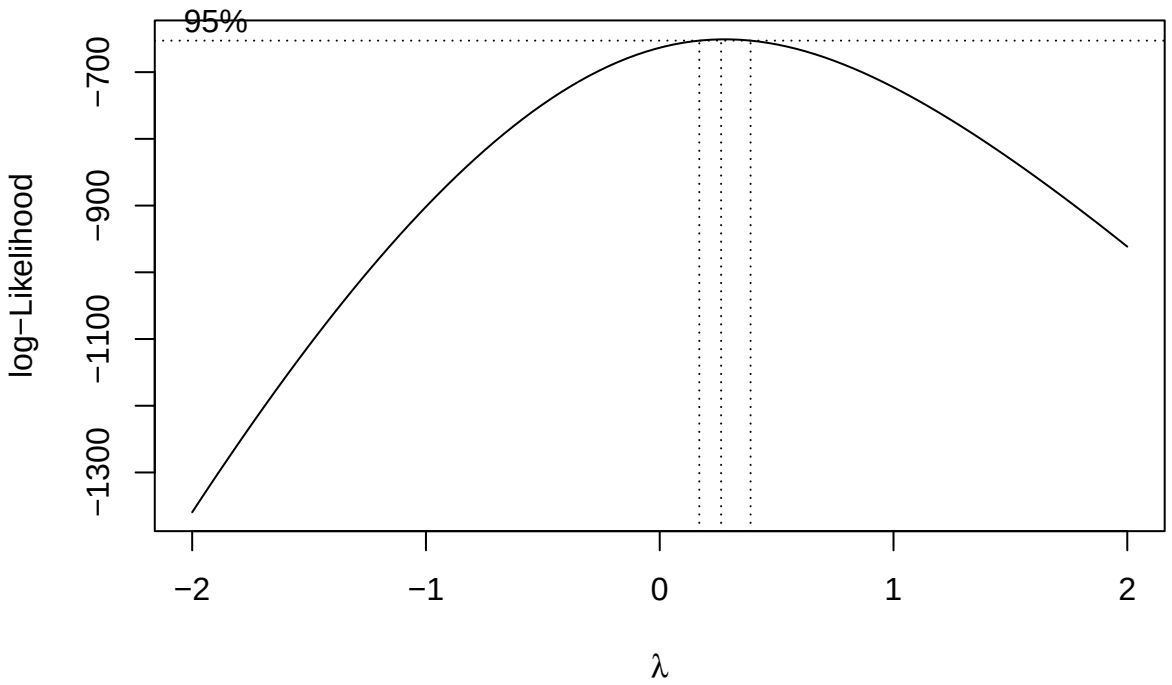
```
par (mfrow = c (1,2))
qqnorm(rstandard(mod), main = "")
abline(0,1, col = "red")
hist (rstandard(mod), main = "", xlab = "Standardized Residuals")
```




```
par (mfrow = c (1,1))
```

Question 2.5 Use the Box-Cox method to find the optimal exponent for a power transformation of the outcome. What is the exponent that you found?

```
library(MASS)
require(MASS)
data(ozone, package = "faraway")
obj <- boxcox(mod, plotit = TRUE)
```




```
obj$x[which.max(obj$y)] # 0.26
```

```
## [1] 0.2626263
```

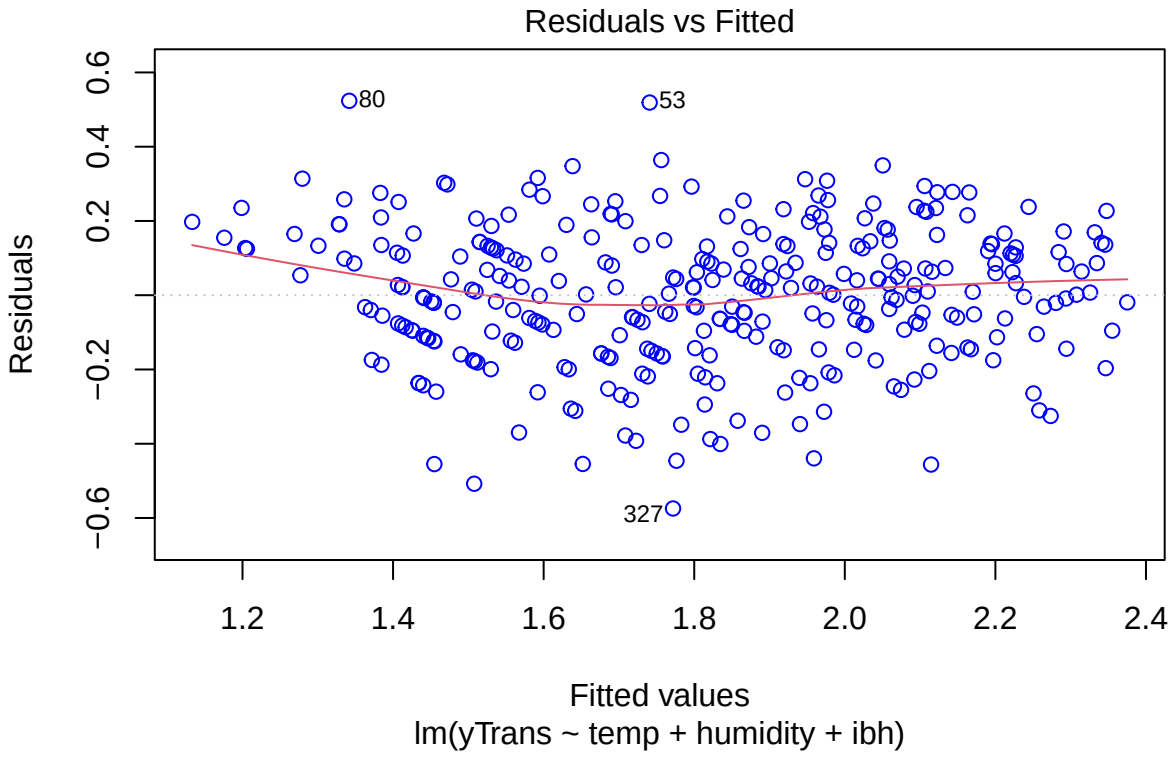
Question 2.6 Create a new variable for the transformed outcome based on the maximum likelihood exponent and fit a new model with the same predictors (including the interaction term). Show the regression output.

```
yTrans <- (ozone$O3)^(0.26)
modTrans <- lm(yTrans ~ temp + humidity + ibh, data = ozone)
summary(modTrans)
```

```
##
## Call:
## lm(formula = yTrans ~ temp + humidity + ibh, data = ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57439 -0.11413  0.00957  0.13218  0.52357
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.000e-01  6.574e-02  13.691  < 2e-16 ***
## temp         1.391e-02  8.579e-04  16.213  < 2e-16 ***
## humidity     3.174e-03  5.447e-04   5.827 1.36e-08 ***
## ibh         -5.148e-05  6.664e-06  -7.725 1.39e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.184 on 326 degrees of freedom
## Multiple R-squared:  0.7156, Adjusted R-squared:  0.713
## F-statistic: 273.4 on 3 and 326 DF,  p-value: < 2.2e-16
```

Question 2.7 Create a residuals vs fitted plot for the new model. Do you see any difference?

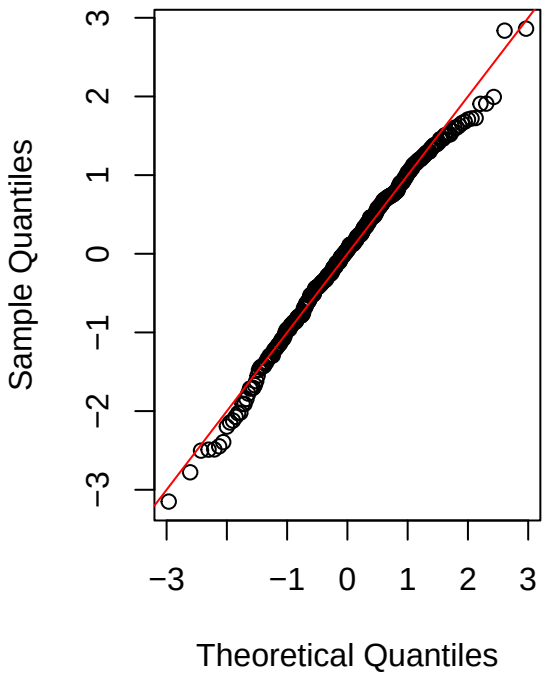
```
plot(modTrans, which = 1, col = "blue")
```

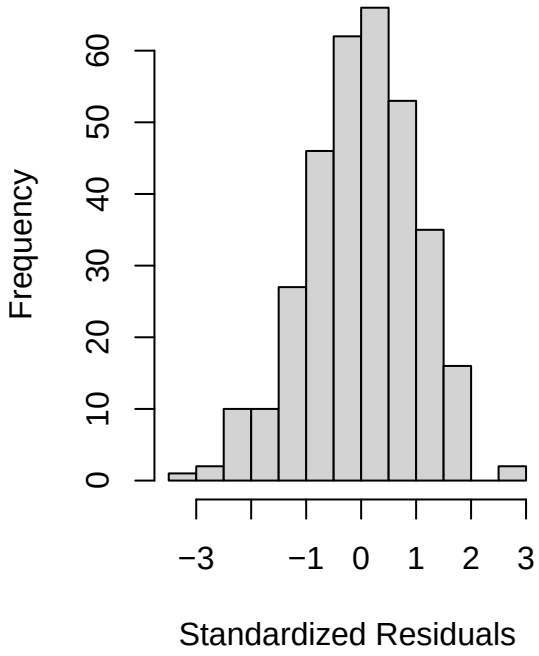



```
# The fitted line is more flat
```

Question 2.8 Create a Q-Q plot and histogram of the standardized residuals for the new model. Do you see any difference?

```
par (mfrow = c (1,2))
qqnorm(rstandard(modTrans), main = "")
abline(0,1, col = "red")
hist (rstandard(modTrans), main = "", xlab = "Standardized Residuals")
```




```
par (mfrow = c (1,1))
```