PSTAT 126 - Assignment 3

Fall 2022

Due: Tuesday, October 18 at 11:59 pm on Canvas

Note: ***Submit both your*** **`Rmd`** ***and generated pdf file to Canvas.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

```
library(alr4)
library(ggplot2)
data('Heights')
```

1. This problem uses the data set Heights from the alr4 package, which contains the heights of n = 1375 pairs of mothers (mheight) and daughters (dheight) in inches.

(a) Compute the regression of dheight on mheight, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the results of these computa- tions.

Solution:

```
x <- Heights$mheight
y <- Heights$dheight
xbar <- mean(x)
ybar <- mean(y)
n <- length(y)
fit <- lm(y~x)
# Regression
Sxx <- sum((x-xbar)^2)
Syy <- sum((y-ybar)^2)
Sxy <- sum((x-xbar)*(y-ybar))

# Estimates
b1 = Sxy/Sxx
b1
```

```
## [1] 0.541747
```

```
b0 = ybar - b1*xbar
b0
```

```
## [1] 29.91744
```

```
yhat <- b0+b1*x

sse=sum((y-yhat)^2)
```

```r
mse=sse/(n-2)

ssr <- sum((yhat-ybar)^2)
sstot <- sum((y-ybar)^2)

# Standard errors
se_b1 <- sqrt(mse/Sxx)
se_b1
```

```
## [1] 0.02596069
```

```r
se_b0 <-sqrt(mse*((1/n)+(xbar^2/Sxx)))
se_b0
```

```
## [1] 1.622469
```

```r
# Coefficient of Determination
r <- ssr/sstot
r
```

```
## [1] 0.2407957
```

```r
# Estimate of Variance
var_est <- summary(fit)$sigma^2
var_est
```

```
## [1] 5.136167
```

```r
# Also
mse
```

```
## [1] 5.136167
```

From the computations above we have that the estimates of β_0 and β_1 are 29.91744 and 0.541747. Also, our standard errors are 1.622469 and 0.02596069. The coefficient of determination is 0.2407957 and the estimate of variance is 5.136167. We know by the coefficient of determination that about 24% of the data fits the regression model, or that about 24% of the variablity of the daughters' height can be explained by the mothers' height.

(b) Obtain a 99% confidence interval for β_1 from the data.

Solution:

```r
confint(fit, level = 0.99) # x = mheight
```

```
##                  0.5 %     99.5 %
## (Intercept) 25.7324151 34.1024585
## x            0.4747836  0.6087104
```

Our 99% confidence interval for β_1 is [0.4747836,0.6097104]

(c) Obtain a predicted value and 90% prediction interval for a daughter whose mother is 58 inches tall.

Solution:

```
predict(fit, data.frame(x=58),interval='predict', level=0.90)
```

```
##        fit      lwr      upr
## 1 61.33876 57.60229 65.07523
```

Our predicted value is 61.33876 and our prediction interval is [57.60229,65.07523]

2. This problem uses the data set prostate from the faraway package (see problem 2 from HW 2).

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

a) Using the variable lpsa as the response and lcavol as the predictor, use R to produce an ANOVA table for this regression fit.

Solution:

```
data('prostate')
lpsa <- prostate$lpsa
lcavol <- prostate$lcavol
```

```
newfit <- lm(lpsa~lcavol)
anova(newfit)
```

```
## Analysis of Variance Table
##
## Response: lpsa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lcavol     1 69.003  69.003  111.27 < 2.2e-16 ***
## Residuals 95 58.915   0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) In the ANOVA table from part a), which quantity represents the variability in lpsa which is left unexplained by the regression?

Solution: The quantity 58.915 represents this variability.

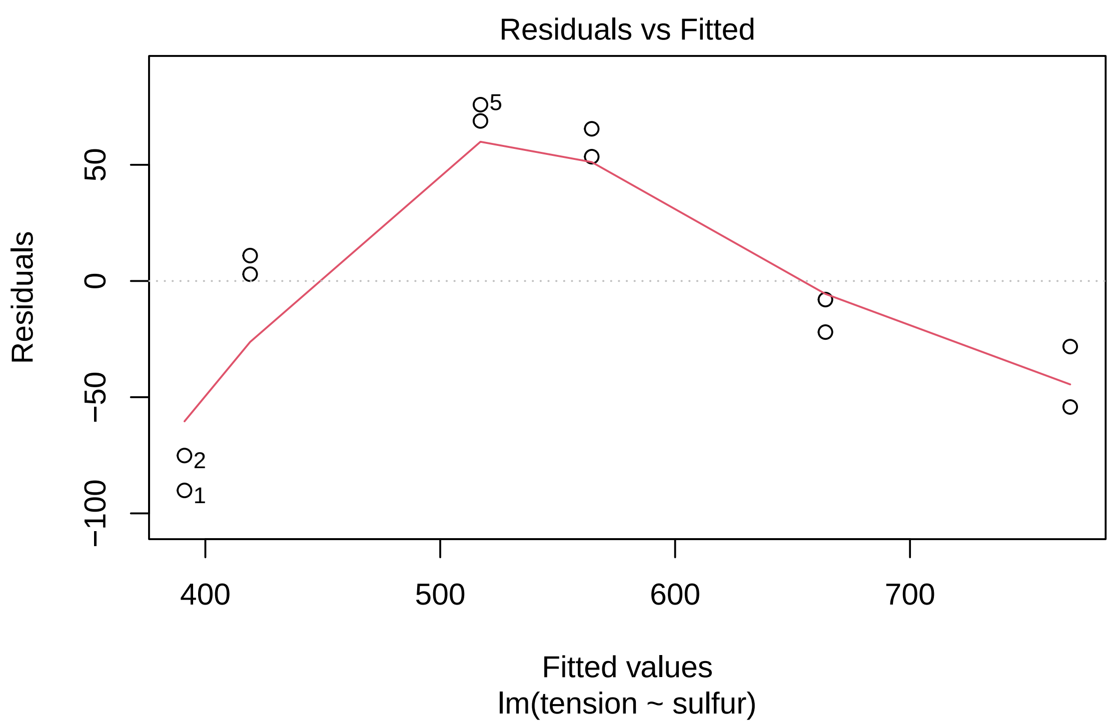
3. This problem uses the data set baeskel from the alr4 package.

```r
library(alr4)
data('baeskel')
```

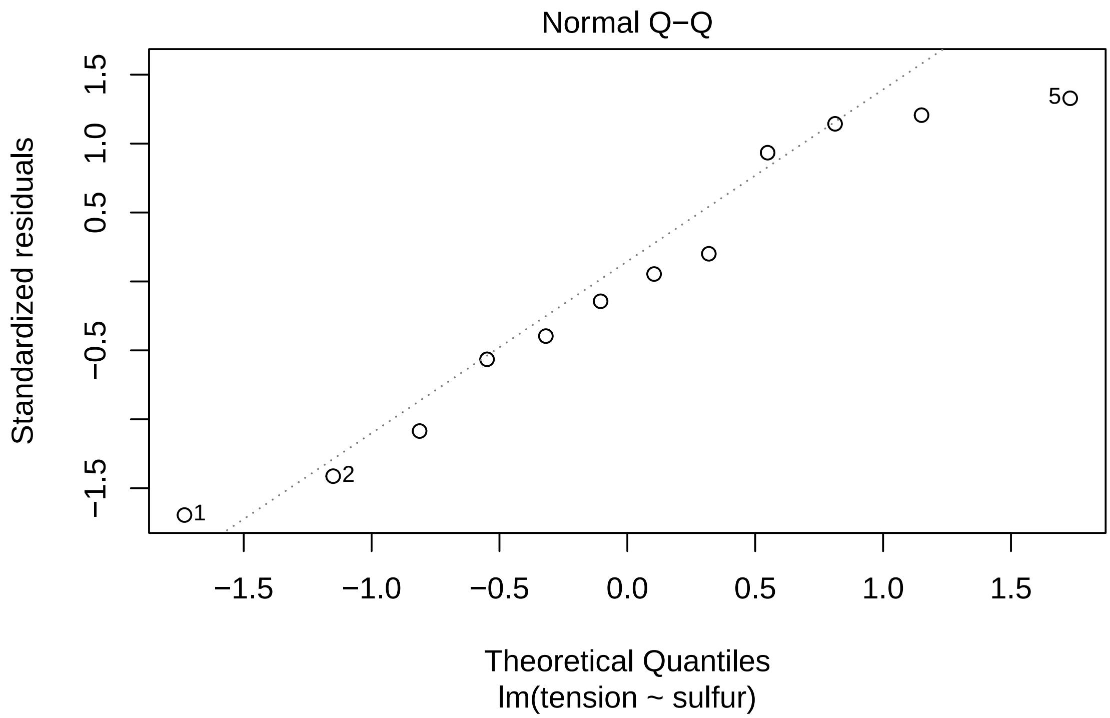
a) Fit the regression model with Tension as response and Sulfur predictor, and produce three diagnostic plots: Residuals vs. Fitted, Scale-Location and a QQ-plot. Comment on any violation of the standard linear model assumptions seen in these plots.

   Solution:

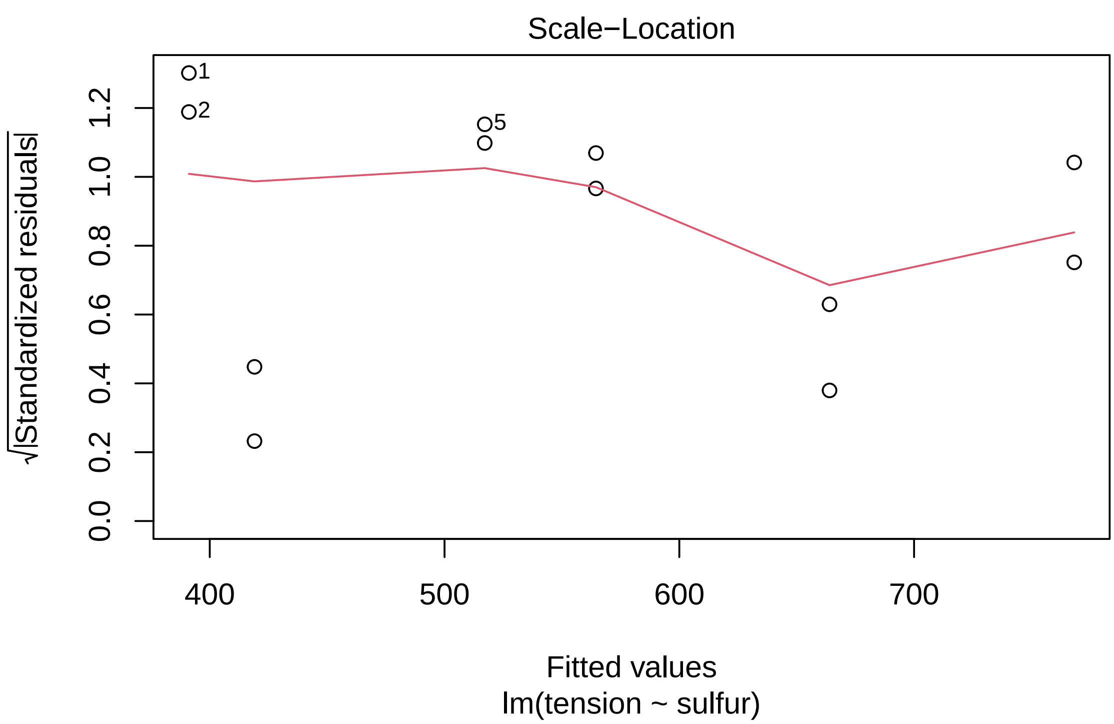
```r
sulfur <- baeskel$Sulfur
tension <- baeskel$Tension
fit3 <- lm(tension~sulfur)
plot(fit3,which = 1)
```




```r
plot(fit3, which = 2)
```




```
plot(fit3, which = 3)
```

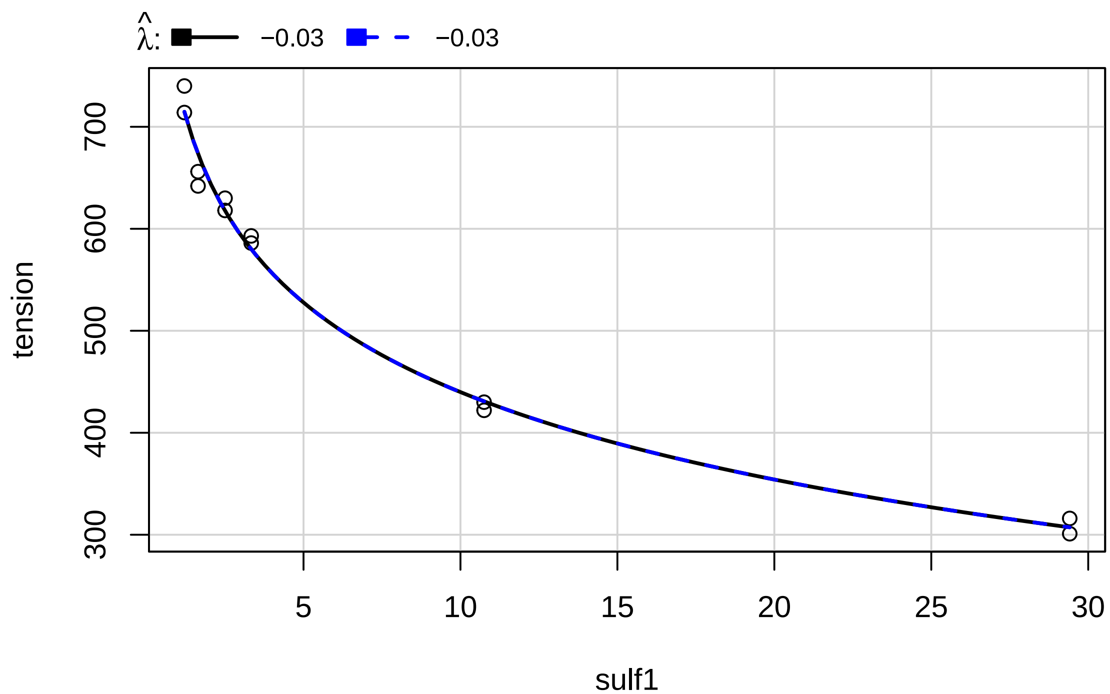


Plot 1- Violation of Linearity Plot 3- Violation of Linearity and Constant error variance

b) Consider two alternative models given by the predictor transformations 1/Sulfur and log(Sulfur): With Sulfur on the horizontal axis and Tension on the vertical axis, fit these two alternatives and plot the regression fits along with the fit from part a). **Note that the two fits from this part will not be linear, since the predictor was transformed.** Hint: The R function invTranPlot is useful here.

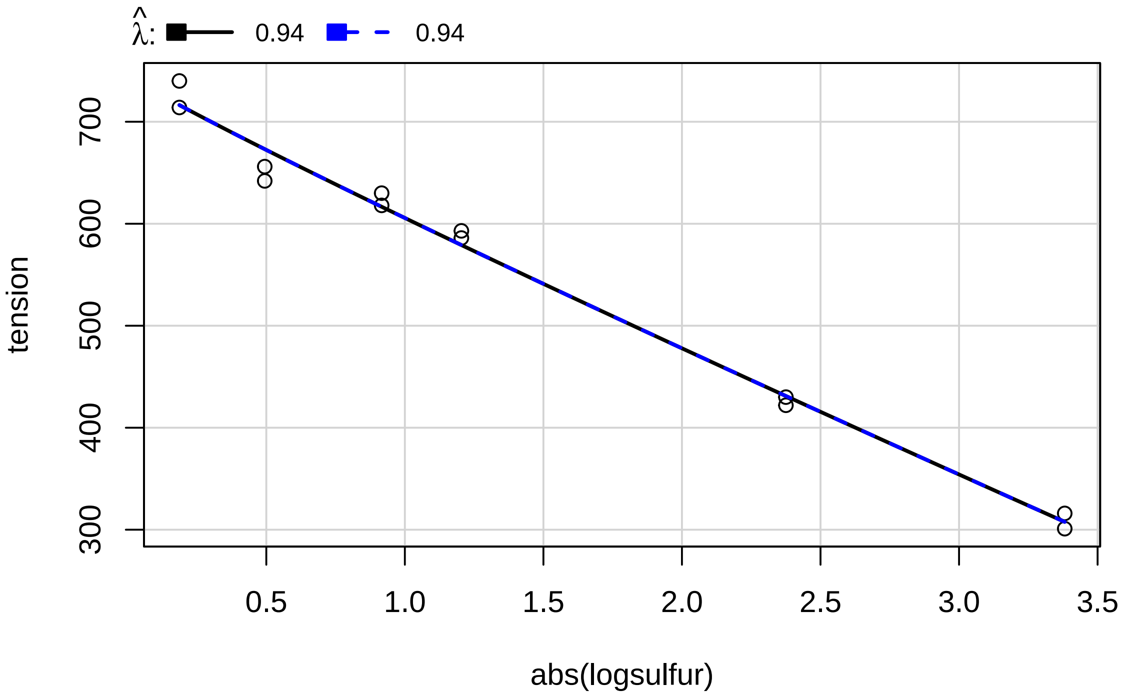
Solution:

```
sulf1 <- 1/sulfur
invTranPlot(tension~sulf1, lambda=invTranEstimate(sulf1,tension)$lambda)
```




```
##      lambda      RSS
## 1 -0.03442 2484.107
## 2 -0.03442 2484.107
```

```
logsulfur <- log(sulfur)
invTranPlot(tension~abs(logsulfur), lambda=invTranEstimate(abs(logsulfur),tension)$lambda)
```




```
##      lambda      RSS
## 1 0.9379937 2446.661
## 2 0.9379937 2446.661
```