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(UN11)
```

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: β_0 is estimated to be 29.91744 with a standard error of 1.62247. β_1 is estimated to be 0.54175 with a standard error of 0.02596. The coefficient of determination, R^2, is 0.2402. The estimate of variance, $\hat{\sigma}^2$, the square of the residual standard error, $\hat{\sigma}^2 = 5.135$. This model shows a significant relationship between the heights of daughters and their mothers, with approximately 24% of the vairability in daughters' heights being explained by their mothers' height.

```
data(Heights)
heights_lm <- lm(dheight~mheight,Heights)
summary(heights_lm)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

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

Solution: 99% CI: $\hat{\beta}_1 \pm t_{n-2,\alpha/2} * SE(\hat{\beta}_1)$

This confidence interval indicates you can be 99% confident that the true value of β_1 lies between 0.4749 and 0.6086.

```
b1_lower = 0.54175 - 0.02596*2.576
b1_upper = 0.54175 + 0.02597*2.576
```

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

Solution:

```
newobs <- data.frame(mheight=c(58))
predict(heights_lm, newdata = newobs, interval = 'predict', level = 0.90)
```

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

```
#predicted value is 61.339, with a 90% prediction interval of (57,602, 65.075)
```

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

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", package = "faraway")
prostate_lm <- lm(lpsa~lcavol, prostate)
prostate_anova <- anova(prostate_lm)
print(prostate_anova)
```

```
## 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?

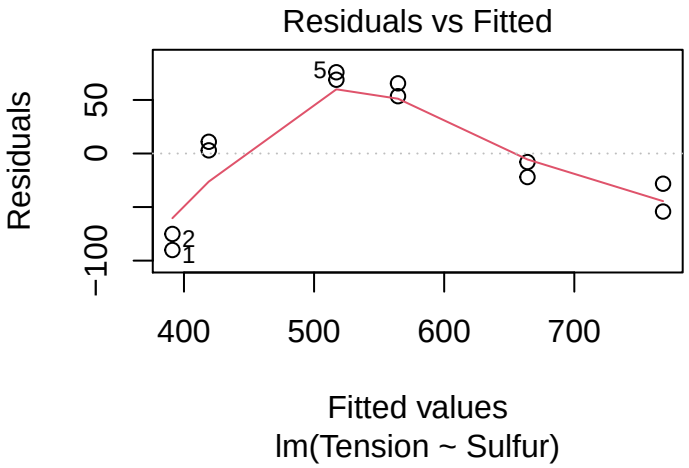
Solution: The Residual Sum of Squares (RSS), which is equal to 58.915 in this casee, represents the variability in lpsa unexplained by the model.

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

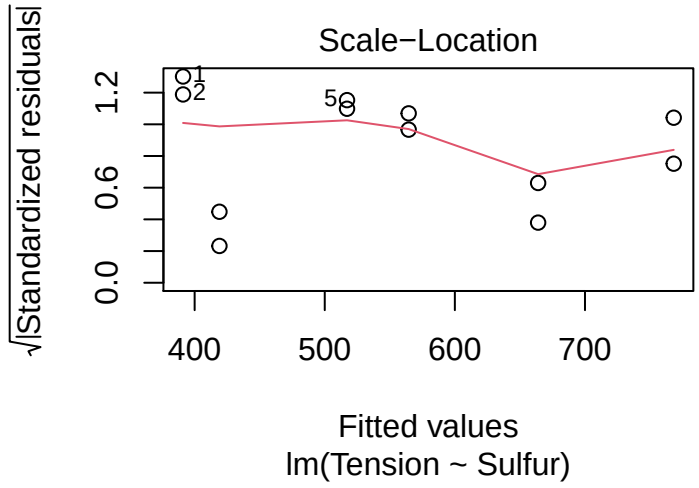
a) Fit the regression model with Tension as response and Sulfur predictor, and produce three diagnostic plots: Residuals vs. Fitted, Scale-Location and a QQ-plot. Comment on any violation of the standard linear model assumptions seen in these plots.

Solution: There are no violations to the standard linear model assumptions. There is a potential skew of the residuals, but not enough data points to make a definitive conclusion.

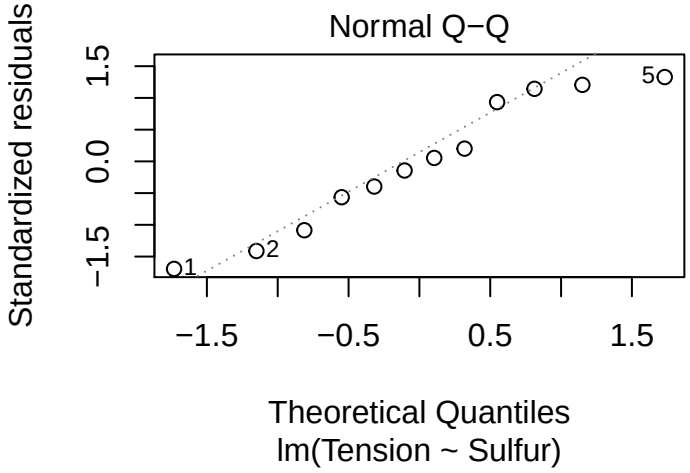
```
data(baeskel)
tension_lm <- lm(Tension ~ Sulfur, baeskel)
plot(tension_lm, which = 1)
```




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



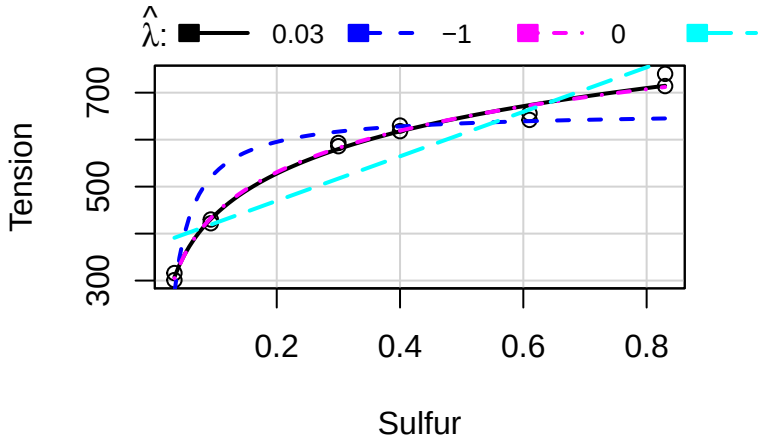

```
plot(tension_lm, which = 2)
```



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:

```
invTranPlot(Tension~Sulfur,data = baeskel, lambda = c(-1,0,1))
```




```
##     lambda       RSS
## 1  0.03442  2484.107
## 2 -1.00000 35691.735
## 3  0.00000  2535.896
## 4  1.00000 35824.332
```