Homework 3

Emily Lu

4/27/2019

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 computations.

The slope estimate is $\beta_0 = 29.91744$ and the intercept estimate is $\beta_1 = 0.541747$. The standard errors for β_0 and β_1, respectively, are 1.622469 and 0.02596069. The coefficient of determination is 0.2407957, meaning that 24.08% of the variability of the daughters' height rate is explained by the model. The estimate of variance is 5.136167.

```
library(alr4)
```

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

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

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data("Heights")

x <- Heights$mheight
y <- Heights$dheight
n <- length(x)
fit <- lm(y~x)

# intercept estimates
b0 <- coef(fit)[1]
b1 <- coef(fit)[2]
coef(fit)
```

```
## (Intercept)           x
##   29.917437    0.541747
```

```
# standard error for b0
sigmahat <- summary(fit)$sigma
Sxx <- sum((x - mean(x))^2)
se_b0 <- sigmahat*sqrt(1/n + mean(x)^2/Sxx)
se_b0
```

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

```
# standard error for b1
se_b1 <- sigmahat/sqrt(Sxx)
se_b1
```

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

```r
# compute Coefficient of Determination
summary(fit)$r.squared
```

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

```r
# estimate of variance
summary(fit)$sigma^2
```

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

(b) Obtain a 90% confidence interval for β_1 from the data.

90% CI for β_1 is (0.4990166, 0.5844774)

```r
confint(fit, level = 0.9)
```

```
##                    5 %       95 %
## (Intercept) 27.2469103 32.5879633
## x            0.4990166  0.5844774
```

(c) Obtain a predicted value and 99% prediction interval for a daughter whose mother is 61 inches tall.

The predicted value is 62.964 and the 99% prediction interval is (57.11531, 68.8127).

```r
new <- data.frame(x = 61)
predict(fit, new, se.fit = TRUE,
        interval = "prediction", level = 0.99,
        type = "response")$fit
```

```
##      fit      lwr     upr
## 1 62.964 57.11531 68.8127
```

2. This problem uses the data set prostate from the faraway package.

```r
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
```

```r
data("prostate")
lpsa <- prostate$lpsa
lcavol <- prostate$lcavol
```

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

```r
x <- prostate$lcavol
y <- prostate$lpsa
fit1 <- lm(y~x)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          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?

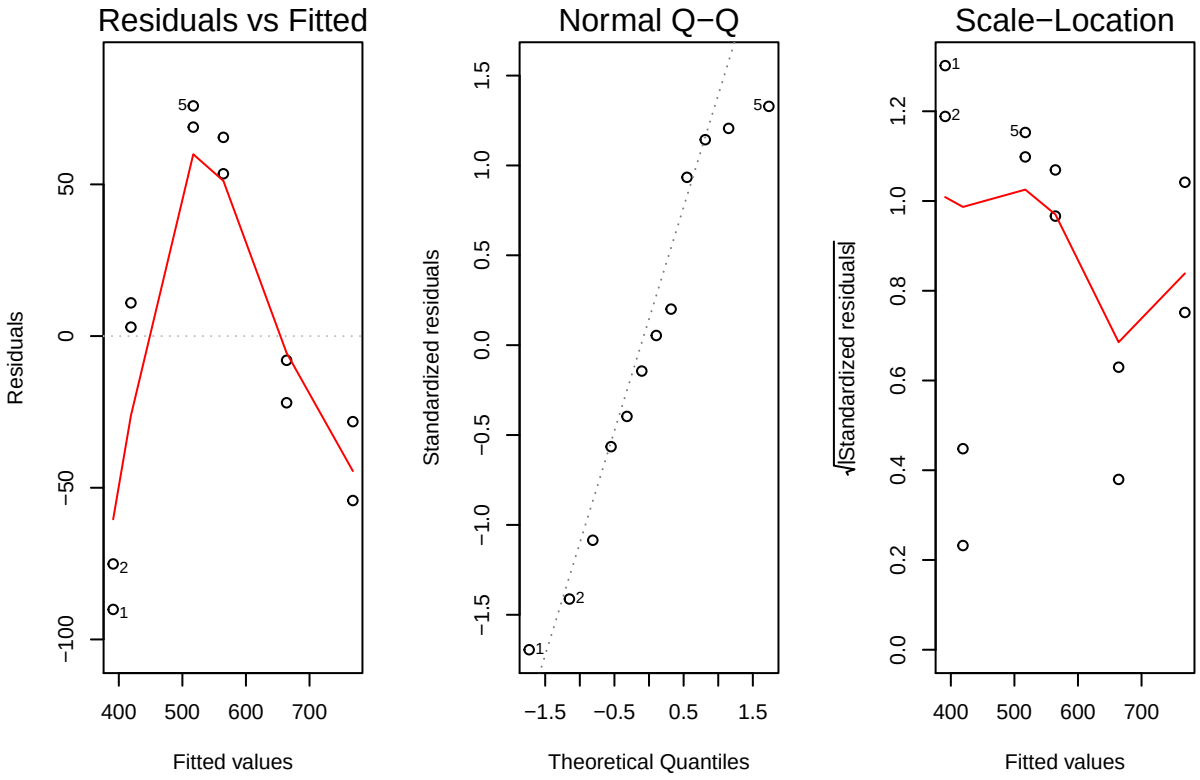
58.915 is the amount of variability in lspa unexplained by the regression.

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.

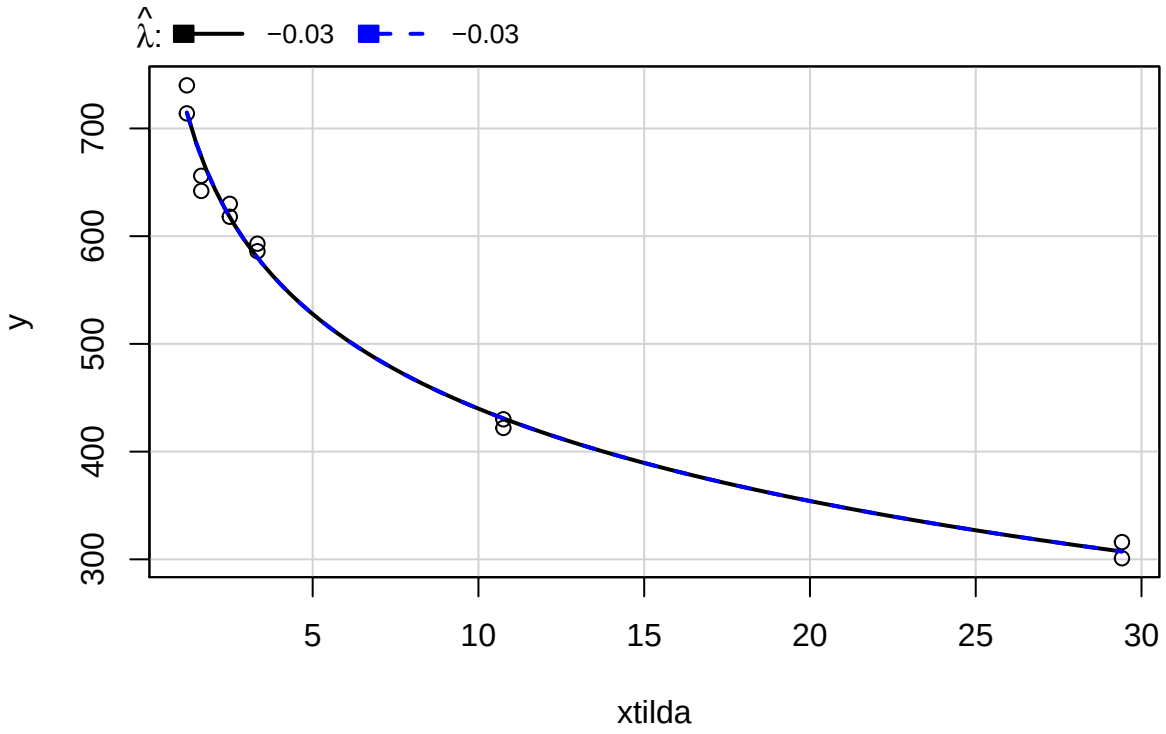
```r
x <- baeskel$Sulfur
y <- baeskel$Tension
fit <- lm(y~x)

par(mfrow = c(1, 3))
for(j in 1:3){
  plot(fit, which = j)}
```



The Residual vs. Fitted plot shows violation of the Linearity. The Scale-Location plot shows the violation of Linearity and Constation (Error) Variance.

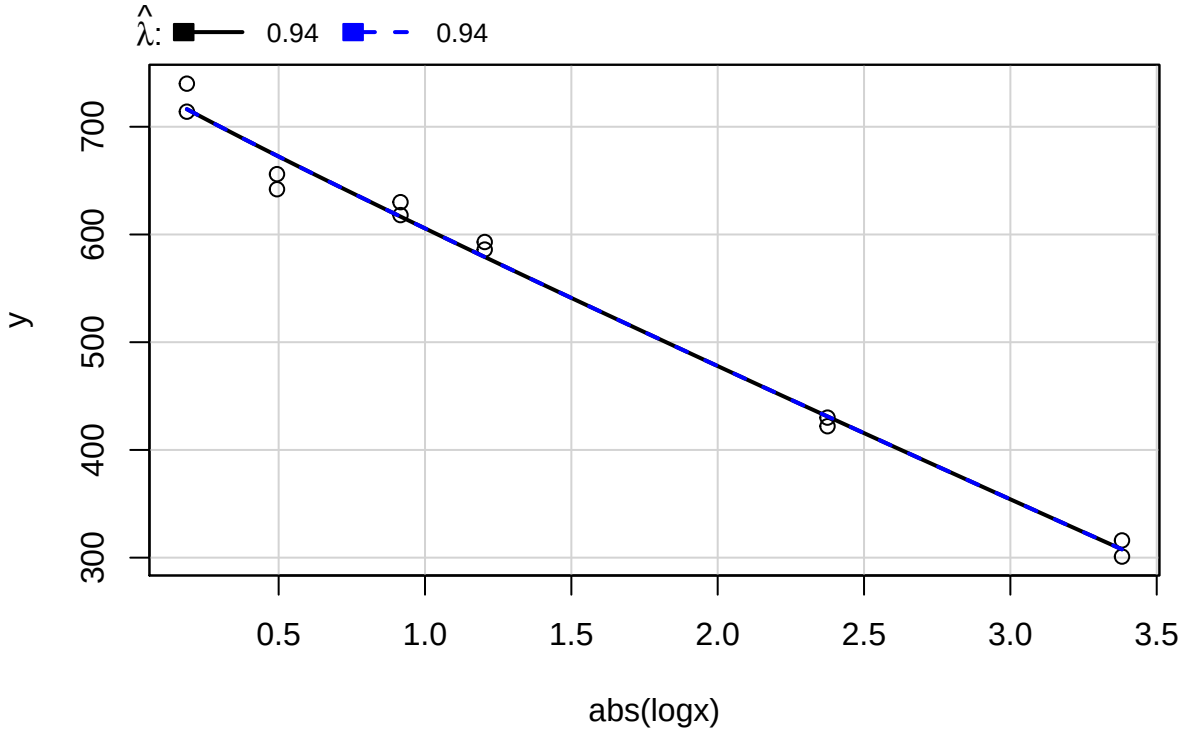
(b) Consider two alternative models given by the predictor transformations 1/Sulfur and log(Sulfur). With Sulfur on the horizontal axis and Tension on the vertical axis, fit these two alternatives and plot the regression fits along with the fit from part a). Note that the two fits from this part will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

```
xtilda <- 1/x
invTranPlot(y~xtilda, lambda = invTranEstimate(xtilda, y)$lambda)
```




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

```
logx <- log(x)
invTranPlot(y~abs(logx), lambda = invTranEstimate(abs(logx), y)$lambda)
```

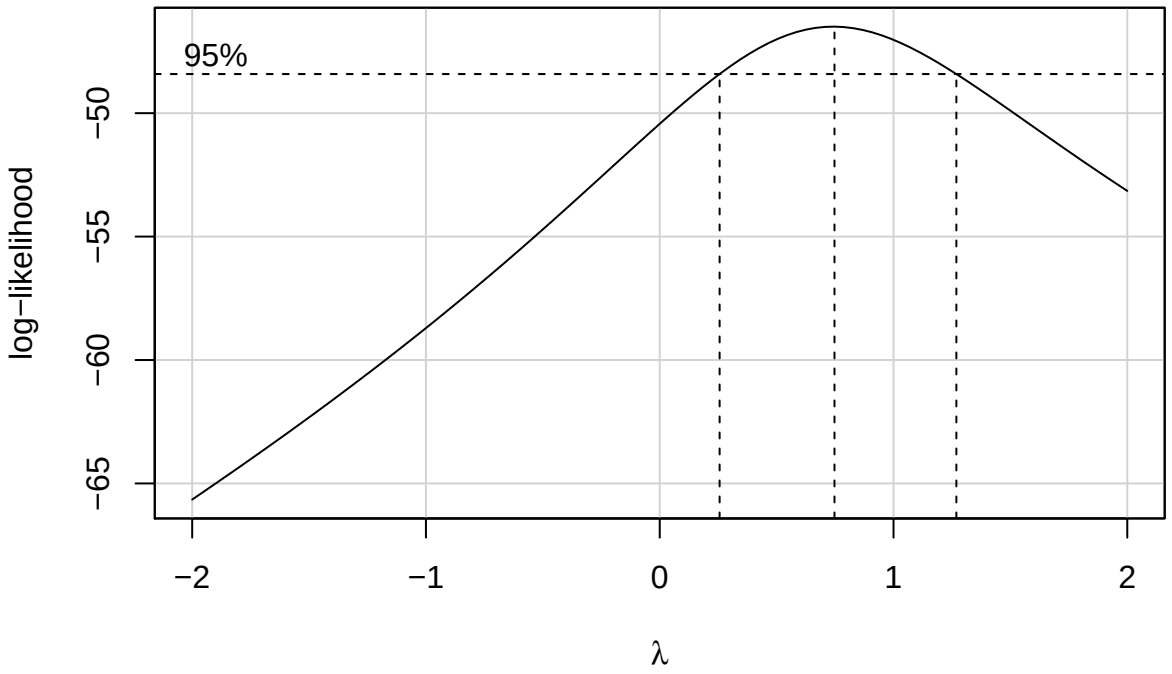



```
##      lambda      RSS
```

```
## 1 0.9379937 2446.661
## 2 0.9379937 2446.661
```

(c) Replace Sulfur by its logarithm, and consider transforming the response Tension. To do this, find and report the optimal power transformation, $\hat{\lambda}_{ML}$ using the BoxCox procedure discussed in class. Should you transform the variable? Explain.

```
fit2 <- lm(y ~ log(x))
bc <- boxCox(fit2)
```




```
lambda.opt <- bc$x[which.max(bc$y)]
lambda.opt
```

```
## [1] 0.7474747
```

```
# We can confirm with the summary function:
summary(powerTransform(y~log(x)))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.7426           1       0.2745       1.2106
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df      pval
## LR test, lambda = (0) 7.872495  1 0.0050192
##
## Likelihood ratio test that no transformation is needed
##                            LRT df    pval
## LR test, lambda = (1) 1.070224  1 0.30089
```

The best optimal power transformation then is 0.74. Since 1 is close to the optimal lambda value (0.74) and within the confidence interval for lambda, we would choose $\lambda = 1$. Thus, no transformation is needed.