PSTAT 126 - Assignment 7

Fall 2022

Kayla Katakis

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.*

1. The data set `mantel` in the `alr4` package has a response Y and three predictors X_1, X_2 and X_3, apply the forward selection and backward elimination algorithms, using AIC as a criterion function. Also, find AIC and BIC for all possible models and compare results. Which appear to be the active regressors?
   Solution: X1 and X2 appear to be the active regressors.

```
library(alr4)
names(mantel)
```

```
## [1] "Y"  "X1" "X2" "X3"
```

```
attach(mantel)
library(leaps)

#AIC forward selection
start <- lm(Y~1,mantel)
end <- lm(Y~., mantel)
step(start, scope = list(lower = start, upper= end),direction = 'forward')
```

```
## Start:  AIC=9.59
## Y ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + X3    1   20.6879  2.1121 -0.3087
## + X1    1    8.6112 14.1888  9.2151
## + X2    1    8.5064 14.2936  9.2519
## <none>              22.8000  9.5866
##
## Step:  AIC=-0.31
## Y ~ X3
##
##        Df Sum of Sq    RSS      AIC
## <none>             2.1121 -0.30875
## + X2    1  0.066328 2.0458  1.53172
## + X1    1  0.064522 2.0476  1.53613
```

```
##
## Call:
## lm(formula = Y ~ X3, data = mantel)
##
## Coefficients:
## (Intercept)           X3
##      0.7975       0.6947
```

```r
#AIC backward selection
step(end, direction = 'backward')
```

```
## Start:  AIC=-285.77
## Y ~ X1 + X2 + X3
##
##        Df Sum of Sq    RSS      AIC
## - X3    1    0.0000 0.0000 -287.749
## <none>              0.0000 -285.768
## - X1    1    2.0458 2.0458    1.532
## - X2    1    2.0476 2.0476    1.536
##
## Step:  AIC=-287.75
## Y ~ X1 + X2
##
##        Df Sum of Sq    RSS      AIC
## <none>               0.000 -287.749
## - X2    1    14.189 14.189    9.215
## - X1    1    14.294 14.294    9.252

##
## Call:
## lm(formula = Y ~ X1 + X2, data = mantel)
##
## Coefficients:
## (Intercept)           X1           X2
##       -1000            1            1
```

```r
#AIC and BIC for all possible models
sub1 <- lm(Y~.,mantel)
sub2 <- lm(Y~X1, mantel)
sub3 <- lm(Y~X2, mantel)
sub4 <- lm(Y~X3, mantel)
sub5 <- lm(Y~X1+X2, mantel)
sub6 <- lm(Y~X2+X3, mantel)
sub7 <- lm(Y~X1+X3, mantel)
sub8 <- lm(Y~1, mantel)

subsets <- list(sub1,sub2,sub3,sub4,sub5,sub6,sub7,sub8)
for (i in subsets){
  print(extractAIC(i))
  print(BIC(i))
}
```

```
## [1]    4.0000 -285.7684
## [1] -271.5318
## [1] 2.000000 9.215066
## [1] 24.23276
## [1] 2.000000 9.251865
## [1] 24.26956
## [1]  2.0000000 -0.3087485
## [1] 14.70895
## [1]    3.0000 -287.7494
## [1] -273.1222
## [1] 3.000000 1.531716
```

```
## [1] 16.15885
## [1] 3.000000 1.536128
## [1] 16.16326
## [1] 1.000000 9.586613
## [1] 24.99487
```

2. In an unweighted regression problem with $n = 54$, $p = 4$, the results included $\hat{\sigma} = 4.0$ and the following statistics for four of the cases:

e_i	h_{ii}
1.000	0.900
1.732	0.750
9.000	0.250
10.295	0.185

For each of these four cases, compute r_i, D_i, and t_i. Test each of the four cases to be an outlier. Make a qualitative statement about the influence of each case on the analysis.
Solution: Cases 3 and 4 are influential outliers due to the small t-test values that allow us to reject the null hypothesis and conclude that these values are outliers.

```r
n<-54
p<-4
sigma_hat <- 4.0
calc_ri <- function(ei,sigma_hat, hii){
  r_i <- ei /(sigma_hat*sqrt(1-hii))
  return(r_i)
}
calc_di <-function(p, r_i, hii){
  d_i <- ((1/p)*(r_i^2))*(hii/(1-hii))
  return(d_i)
}
calc_ti<- function(ei,p,n,sigma_hat,hii){
  s_i2 <- (((n-p)*sqrt(sigma_hat)-(ei^2/(1-hii))/(n-p-1)))
  t_i <- ei/sqrt(s_i2*(1-hii))
  return(t_i)
}
#first row values
calc_ri(1.000,sigma_hat,0.900)
```

```
## [1] 0.7905694
```

```r
calc_di(p,calc_ri(1.000,sigma_hat,0.900), 0.900)
```

```
## [1] 1.40625
```

```r
calc_ti(1.000,p, n, sigma_hat, 0.900)
```

```
## [1] 0.3165509
```

```r
print(" ")
```

```
## [1] " "
```

```r
#second row values
calc_ri(1.732,sigma_hat,0.750)
```

```
## [1] 0.866
```

```r
calc_di(p,calc_ri(1.732,sigma_hat,0.750), 0.750)
```

```
## [1] 0.562467
```

```r
calc_ti(1.732,p, n, sigma_hat, 0.750)
```

```
## [1] 0.3468249
```

```r
print(" ")
```

```
## [1] " "
```

```r
#third row values
calc_ri(9.000,sigma_hat,0.250)
```

```
## [1] 2.598076
```

```r
calc_di(p,calc_ri(9.000,sigma_hat,0.250), 0.250)
```

```
## [1] 0.5625
```

```r
calc_ti(9.000,p, n, sigma_hat, 0.250)
```

```
## [1] 1.050876
```

```r
print(" ")
```

```
## [1] " "
```

```r
#fourth row values
calc_ri(10.295,sigma_hat,0.185)
```

```
## [1] 2.850937
```

```r
calc_di(p,calc_ri(10.295,sigma_hat,0.185), 0.185)
```

```
## [1] 0.4612424
```

```r
calc_ti(10.295,p, n, sigma_hat, 0.185)
```

```
## [1] 1.155815
```

```r
print(" ")
```

```
## [1] " "
```

```r
#testing t values for outliers
dt(0.3165509,49)
```

```
## [1] 0.3771498
```

```r
dt(0.3468249,49)
```

```
## [1] 0.3733137
```

```r
dt(1.050876,49)
```

```
## [1] 0.2273593
```

```r
dt(1.155815,49)
```

```
## [1] 0.2026043
```

3. The `lathe1` data set from the `alr4` package contains the results of an experiment on characterizing the life of a drill bit in cutting steel on a lathe. Two factors were varied in the experiment, `Speed` and `Feed`

rate. The response is `Life`, the total time until the drill bit fails, in minutes. The values of `Speed` and `Feed` in the data have been coded by computing

$$\texttt{Speed} = \frac{\text{Actual speed in feet per minute} - 900}{300}$$

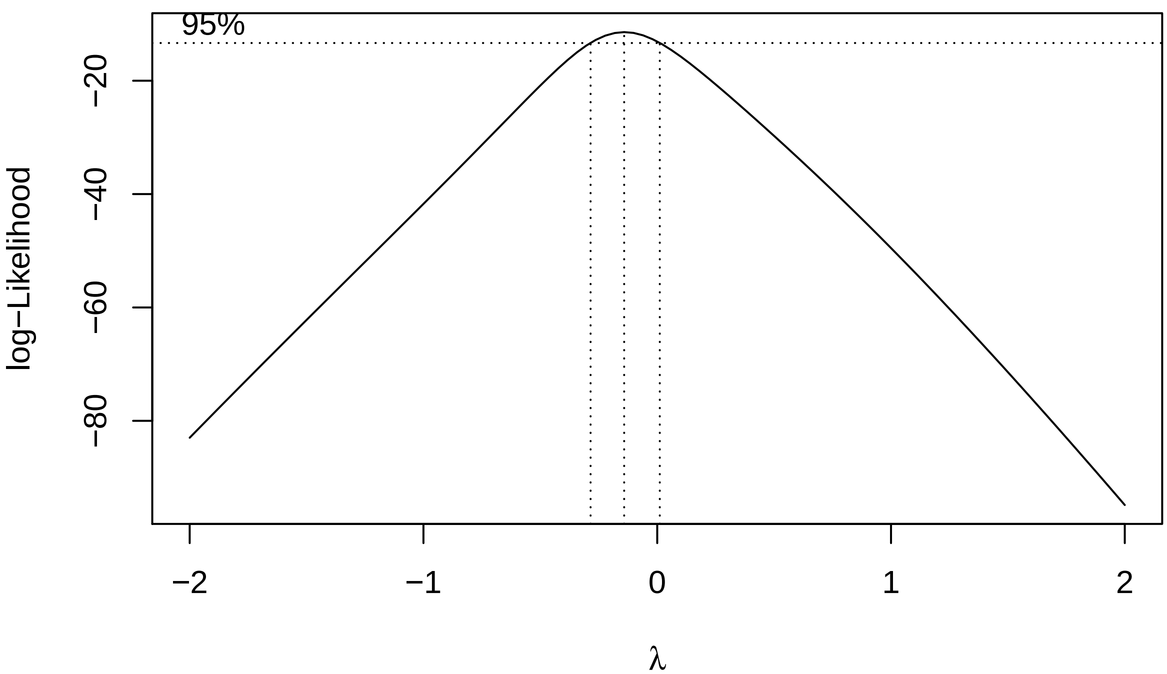
$$\texttt{Feed} = \frac{\text{Actual feed rate in thousandths of an inch per revolution} - 13}{6}.$$

(a) Starting with the full second-order model

$$E(\texttt{Life}|\texttt{Speed, Feed}) = \beta_0 + \beta_1\texttt{Speed} + \beta_2\texttt{Feed} + \beta_{11}\texttt{Speed}^2 + \beta_{22}\texttt{Feed}^2 + \beta_{12}\texttt{Speed} * \texttt{Feed},$$

use the Box–Cox method to show that an appropriate scale for the response is the logarithmic scale.
Solution: The 95% confidence interval for the BoxCox transformation contains $\lambda = 0$, so a logarithmic scale would be appropriate.

```
library(MASS)
attach(lathe1)
#names(lathe1)
model2 <- lm(Life~1+Speed+Feed+Speed^2+Feed^2+Speed*Feed, lathe1)
boxcox <- boxcox(model2)
```



(b) Find the two cases that are most influential in the fit of the quadratic mean function for log(`Life`), and explain why they are influential. Delete these points from the data, refit the quadratic mean function, and compare with the fit with all the data.
Solution: Cases 7 and 14 are the most influential in the fit of the quadratic mean because all diagnostic plots reveal that these points do not follow the same trends, thus heavily skewing the model. By removing these points, we get a much more accurate accurate model. The variability explained by the model increases and $p - values$ for predictors Speed and Feed become more significant among other signs of increased accuracy.

```
library(alr4)
data(lathe1)
attach(lathe1)
quad_mean_func <-lm(log(Life)~1+Speed+Feed+Speed^2+Feed^2+Speed*Feed, lathe1)
#plot(quad_mean_func)
```

```r
lathe1_copy <- lathe1[-c(7,14),]
View(lathe1_copy)
quad_mean_func2 <- lm(log(Life)~1+Speed+Feed+Speed^2+Feed^2+Speed*Feed, lathe1_copy)
summary(quad_mean_func)
```

```
##
## Call:
## lm(formula = log(Life) ~ 1 + Speed + Feed + Speed^2 + Feed^2 +
##     Speed * Feed, data = lathe1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82354 -0.30087  0.03213  0.27128  0.76259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.61200    0.10754  14.990 7.72e-11 ***
## Speed       -1.58902    0.13884 -11.445 4.07e-09 ***
## Feed        -0.79023    0.13884  -5.692 3.34e-05 ***
## Speed:Feed  -0.07286    0.17003  -0.428    0.674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4809 on 16 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.8942
## F-statistic: 54.52 on 3 and 16 DF,  p-value: 1.273e-08
```

```r
summary(quad_mean_func2)
```

```
##
## Call:
## lm(formula = log(Life) ~ 1 + Speed + Feed + Speed^2 + Feed^2 +
##     Speed * Feed, data = lathe1_copy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57968 -0.26148 -0.03808  0.22086  0.67589
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6093     0.1005  16.020 2.13e-10 ***
## Speed        -1.6619     0.1291 -12.871 3.79e-09 ***
## Feed         -0.8631     0.1291  -6.685 1.04e-05 ***
## Speed:Feed   -0.1822     0.1625  -1.121    0.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4213 on 14 degrees of freedom
## Multiple R-squared:  0.933,  Adjusted R-squared:  0.9187
## F-statistic:    65 on 3 and 14 DF,  p-value: 1.846e-08
```