PSTAT 126 - Assignment 7

Fall 2022

Due: Tuesday, November 22 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.*

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:

```r
library(alr4)
data("mantel")

mlm <- lm(Y ~ ., data = mantel)
m0 <- lm(Y ~ 1, data = mantel)
l <- length(mantel$Y)
```

```r
# AIC
step(m0, scope = list(lower = m0, upper = mlm),
     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
# BIC
step(m0, scope = list(lower = m0, upper = mlm),
     direction = "forward", k = log(l), trace=0)
```

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

```r
# AIC
step(mlm, scope = list(lower = m0, upper = mlm),
     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
# BIC
step(mlm, scope = list(lower = m0, upper = mlm),
     direction = "backward", k = log(l), trace=0)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = mantel)
```

```
##
## Coefficients:
## (Intercept)           X1           X2
##       -1000            1            1
```

For forward selection for both AIC and BIC, only X3 is an active aggressor. For backward selection for both AIC and BIC, X1 and X2 are active aggressors.

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:

```r
ei <- c(1,1.732,9,10.295)
hii <- c(.9,.75,.25,.185)

ri <- ei[1]/(4*sqrt(1-hii[1]))
Di <- ri[1]^2*hii[1]/4*1/(1-hii[1])
ti <- ri[1]*sqrt((49)/(50-ri[1]^2))
```

```r
for(i in c(2:4)){
  r <- ei[i]/(4*sqrt(1-hii[i]))
  ri <- c(ri,r)

  D <- ri[i]^2*hii[i]/4*1/(1-hii[i])
  Di <- c(Di,D)

  t <- ri[i]*sqrt((49)/(50-ri[i]^2))
  ti <- c(ti,t)

}
```

```r
# For each of the four cases
ri[1:4]
```

```
## [1] 0.7905694 0.8660000 2.5980762 2.8509366
```

```r
Di[1:4]
```

```
## [1] 1.4062500 0.5624670 0.5625000 0.4612424
```

```r
ti[1:4]
```

```
## [1] 0.7875615 0.8637988 2.7653931 3.0840606
```

Notice that for r3, r4 and t3, t4 the values are > 2. Hence these are potential outliers at 3 and 4. Moreover, observing Cook's distance measure Di, notice the 1st case is likely influential, as Cook's distance measures the changes to the fitted values when the i-th observation is removed.

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

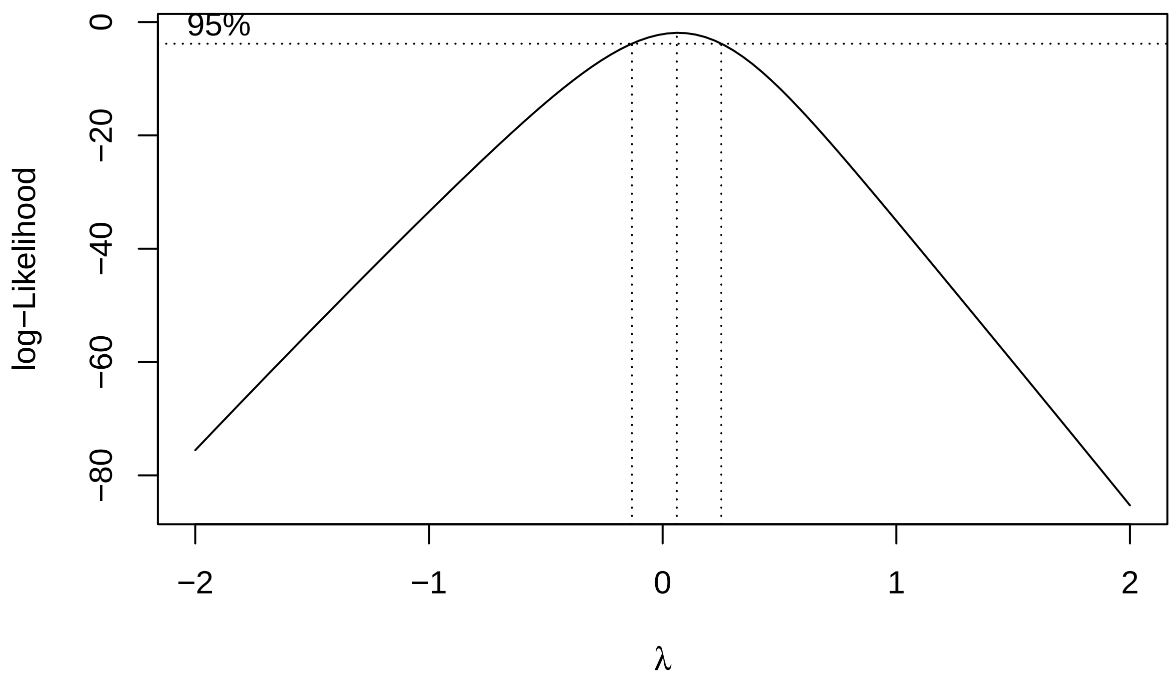
$$\texttt{Speed} = \frac{\text{Actual speed in feet per minute} - 900}{300}$$
$$\texttt{Feed} = \frac{\text{Actual feed rate in thousandths of an inch per revolution} - 13}{6}.$$

(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**:

```r
library(MASS)
library(alr4)
data("lathe1")
lathelm <- lm(Life ~ Speed + Feed +I(Speed^2) + I(Feed^2) + Speed*Feed, data = lathe1)
boxcox(lathelm)
```



lambda = 0 is in the 95% confidence interval, therefore an appropriate scale for the response is the logarithmic scale.

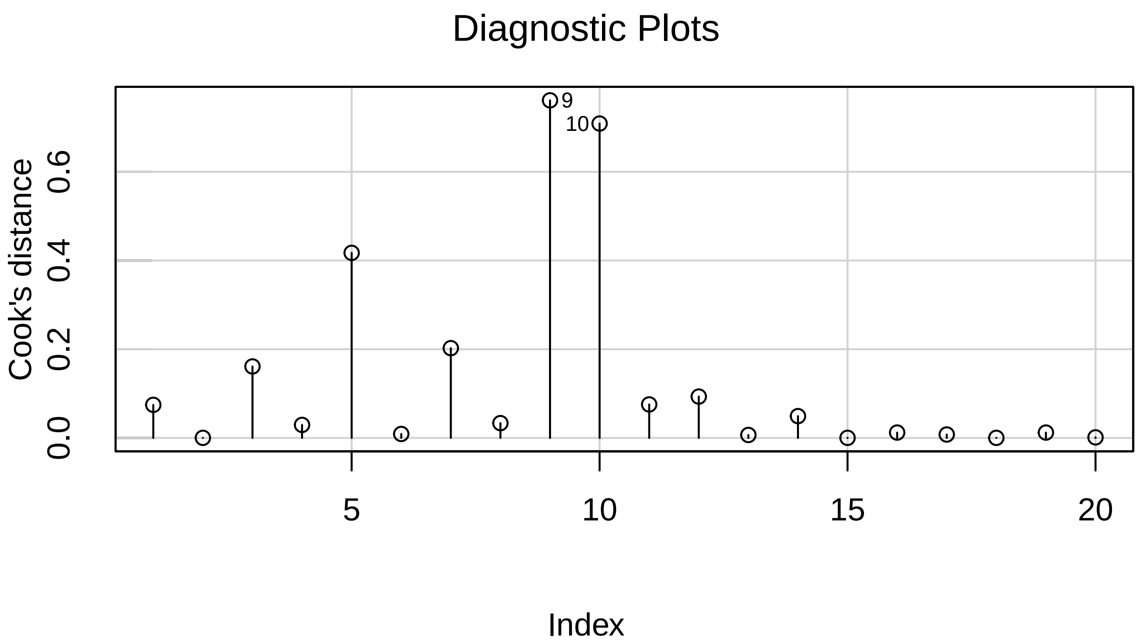
(b) Find the two cases that are most influential in the fit of the quadratic mean function for log(`Life`), and explain why they are influential. Delete these points from the data, refit the quadratic mean function, and compare with the fit with all the data.
Solution:

```
fit <- lm(log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed*Feed, data = lathe1)

cooks <- cooks.distance(fit)
which(cooks > 4/(length(lathe1$Life)-2-1))
```

```
##  5  9 10
##  5  9 10
```

```
influenceIndexPlot(fit,vars = 'Cook',id=list(location = "avoid",n=2, cex = 0.7))
```




```
cds <- cooks.distance(fit)
cds
```

```
##            1            2            3            4            5            6
## 0.0745581876 0.0002358999 0.1611290980 0.0293444172 0.4172638143 0.0089104068
##            7            8            9           10           11           12
## 0.2024479551 0.0333705363 0.7611370235 0.7088115474 0.0755462115 0.0932562838
##           13           14           15           16           17           18
## 0.0066483194 0.0491977930 0.0001916341 0.0121013330 0.0077362334 0.0001916341
##           19           20
## 0.0121013330 0.0012883357
```

We can see that cases 9 and 10 are the most influential. They both are greater than 0.5 and outlies, suggesting that the two cases are likely influential.

```
fit2 <- lm(log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed*Feed, data = lathe1[-c(9,10),])

summary(fit)
```

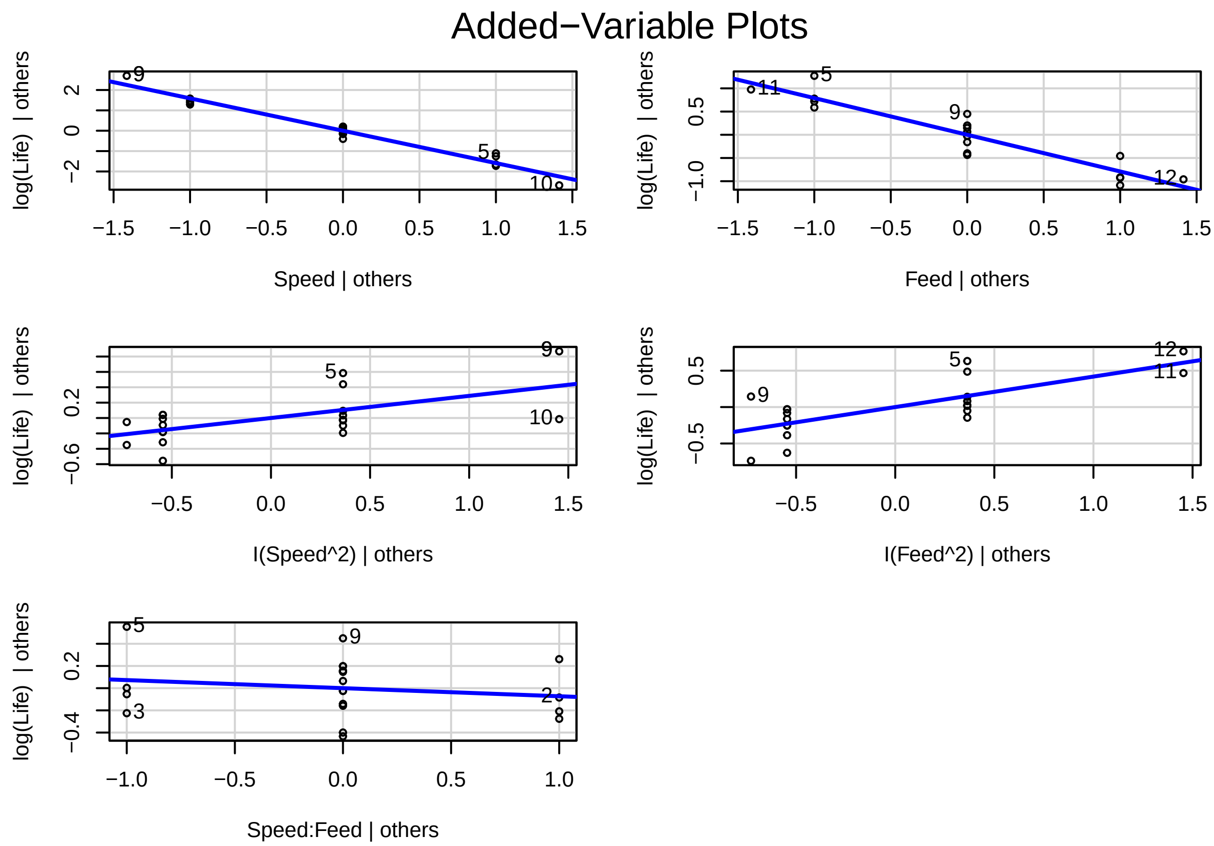
```
##
## Call:
## lm(formula = log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) +
##     Speed * Feed, data = lathe1)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.43349 -0.14576 -0.02494  0.16748  0.47992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18809    0.10508  11.307 2.00e-08 ***
## Speed       -1.58902    0.08580 -18.520 3.04e-11 ***
## Feed        -0.79023    0.08580  -9.210 2.56e-07 ***
## I(Speed^2)   0.28808    0.10063   2.863 0.012529 *
## I(Feed^2)    0.41851    0.10063   4.159 0.000964 ***
## Speed:Feed  -0.07286    0.10508  -0.693 0.499426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2972 on 14 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9596
## F-statistic: 91.24 on 5 and 14 DF,  p-value: 3.551e-10
```

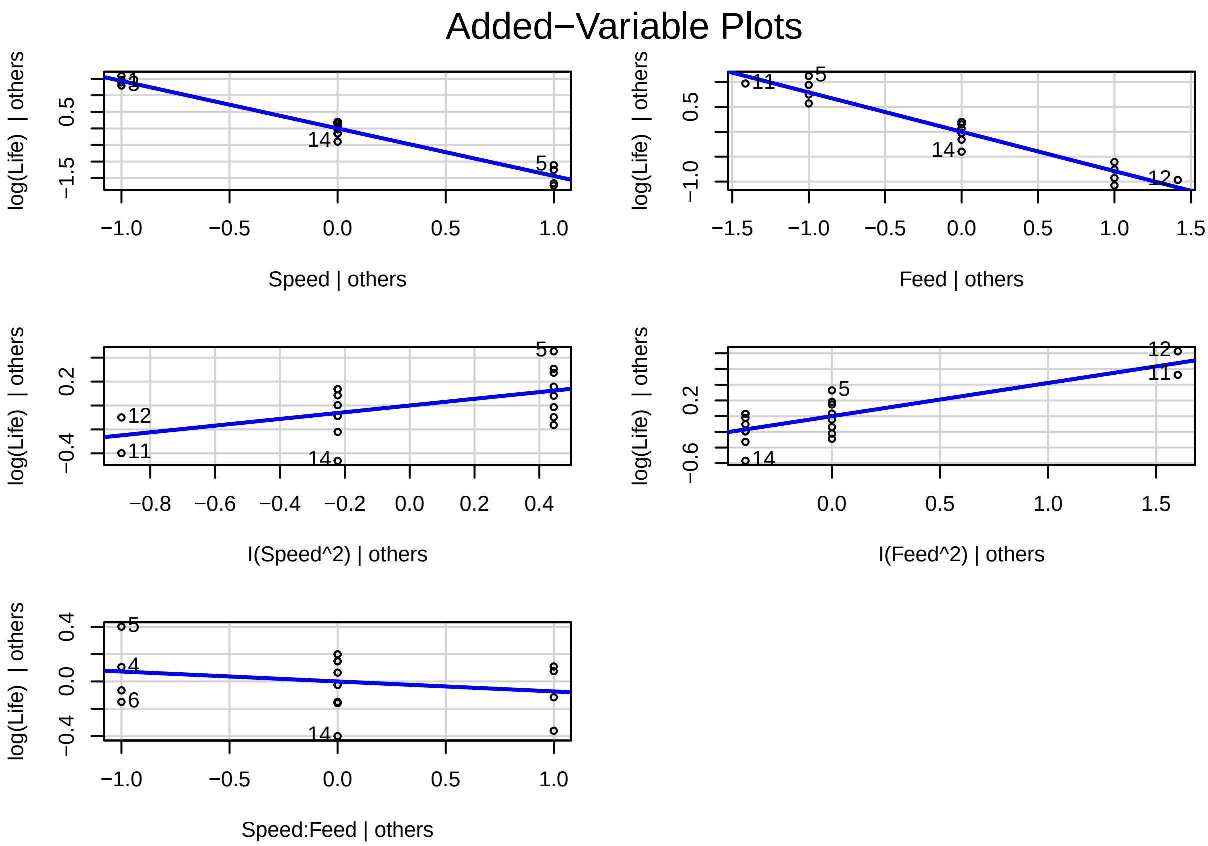
```
summary(fit2)
```

```
##
## Call:
## lm(formula = log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) +
##     Speed * Feed, data = lathe1[-c(9, 10), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39963 -0.14660  0.00387  0.14917  0.32783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18809    0.08241  14.417 6.11e-09 ***
## Speed       -1.43300    0.08241 -17.388 7.10e-10 ***
## Feed        -0.79023    0.06729 -11.743 6.15e-08 ***
## I(Speed^2)   0.28022    0.12363   2.267 0.042700 *
## I(Feed^2)    0.42244    0.09217   4.583 0.000629 ***
## Speed:Feed  -0.07286    0.08241  -0.884 0.394025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2331 on 12 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9658
## F-statistic: 97.07 on 5 and 12 DF,  p-value: 2.804e-09
```

```
avPlots(fit)
```




```
avPlots(fit2)
```



From the summaries and avPlots, we see that the standard errors are uniformly smaller using the reduced data set and that R^2 and adjusted R^2 are larger.