Homework 7

Seth Marceno

6/6/2019

Question 1

Here I read in the data and do forwards AIC selection:

```
data(mantel)
X1 <- mantel$X1
X2 <- mantel$X2
X3 <- mantel$X3
Y <- mantel$Y
fullModel <- ~ X1 + X2 + X3
baseModel <- lm(Y ~ X1)
forward <- step(baseModel, fullModel, direction = 'forward')
```

```
## Start:  AIC=9.22
## Y ~ X1
##
##        Df Sum of Sq     RSS      AIC
## + X2    1    14.189  0.0000 -287.749
## + X3    1    12.141  2.0476    1.536
## <none>               14.1888    9.215
##
## Step:  AIC=-287.75
## Y ~ X1 + X2
##
##        Df  Sum of Sq        RSS     AIC
## <none>                1.5284e-25 -287.75
## + X3    1 5.7989e-28 1.5225e-25 -285.77
```

Here we do backwards AIC selection:

```
m1 <- update(baseModel, fullModel)
backwards <- step(m1, scope = c(lower = ~ X1) , 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
## - 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
```

Here I do forwards BIC selection:

```r
forwardBIC <- step(baseModel, fullModel, direction = 'forward', k = log(length(X1)))
```

```
## Start:  AIC=8.43
## Y ~ X1
##
##        Df Sum of Sq     RSS      AIC
## + X2    1    14.189  0.0000 -288.921
## + X3    1    12.141  2.0476    0.364
## <none>              14.1888    8.434
##
## Step:  AIC=-288.92
## Y ~ X1 + X2
##
##        Df  Sum of Sq        RSS     AIC
## <none>               1.5284e-25 -288.92
## + X3    1 5.7989e-28 1.5225e-25 -287.33
```

Here I do backwards BIC selection:

```r
backwardBIC <- step(m1, scope = c(lower = ~ X1), direction = 'backward', k = log(length(X1)))
```

```
## Start:  AIC=-287.33
## Y ~ X1 + X2 + X3
##
##        Df Sum of Sq    RSS      AIC
## - X3    1    0.0000 0.0000 -288.921
## <none>              0.0000 -287.331
## - X2    1    2.0476 2.0476    0.364
##
## Step:  AIC=-288.92
## Y ~ X1 + X2
##
##        Df Sum of Sq    RSS      AIC
## <none>               0.000 -288.921
## - X2    1    14.189 14.189    8.434
```

Based off of both forwards and backwards selection using BIC and AIC, we see that the active regressors are X1 and X2.

Question 2

Case 1:

```r
e <- 1
hii <- .9
p <- 4
n <- 54
sigma <- 4


ri_1 <- e/ (sigma*(sqrt(1-hii)))
ri_1
```

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

```r
Di_1 <- (1/p)*(ri_1^2)*(hii/(1-hii))
Di_1
```

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

```r
ti_1 <- ri_1*(((n-p-1)/(n-p-(ri_1^2)))^0.5)
ti_1
```

```
## [1] 0.7875615
```

```r
qt(p = 0.975, df = 49, lower.tail = TRUE)
```

```
## [1] 2.009575
```

At level alpha = 0.05, our critical value is 2.01. Since ti_1 is not greater than our critical value, we determine that this point is not influential.

Case 2:

```r
e <- 1.732
hii <- 0.75
p <- 4
n <- 54
sigma <- 4


ri_2 <- e/ (sigma*(sqrt(1-hii)))
ri_2
```

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

```r
Di_2 <- (1/p)*(ri_2^2)*(hii/(1-hii))
Di_2
```

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

```r
ti_2 <- ri_2*(((n-p-1)/(n-p-(ri_2^2)))^0.5)
ti_2
```

```
## [1] 0.8637988
```

Comparing our test statistic to the critical value we found above, again we see our test statistic is smaller, so we conlcude that this point is not influential.

Case 3:

```r
e <- 9
hii <- 0.25
p <- 4
n <- 54
sigma <- 4
```

```r
ri_3 <- e/ (sigma*(sqrt(1-hii)))
ri_3
```

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

```r
Di_3 <- (1/p)*(ri_3^2)*(hii/(1-hii))
Di_3
```

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

```r
ti_3 <- ri_3*(((n-p-1)/(n-p-(ri_3^2)))^0.5)
ti_3
```

```
## [1] 2.765393
```

Since our test statistic is greater than our critical value (2.01) we determine that this point is influential.

Case 4

```r
e <- 10.295
hii <- 0.185
p <- 4
n <- 54
sigma <- 4


ri_4 <- e/ (sigma*(sqrt(1-hii)))
ri_4
```

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

```r
Di_4 <- (1/p)*(ri_4^2)*(hii/(1-hii))
Di_4
```

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

```r
ti_4 <- ri_4*(((n-p-1)/(n-p-(ri_4^2)))^0.5)
ti_4
```

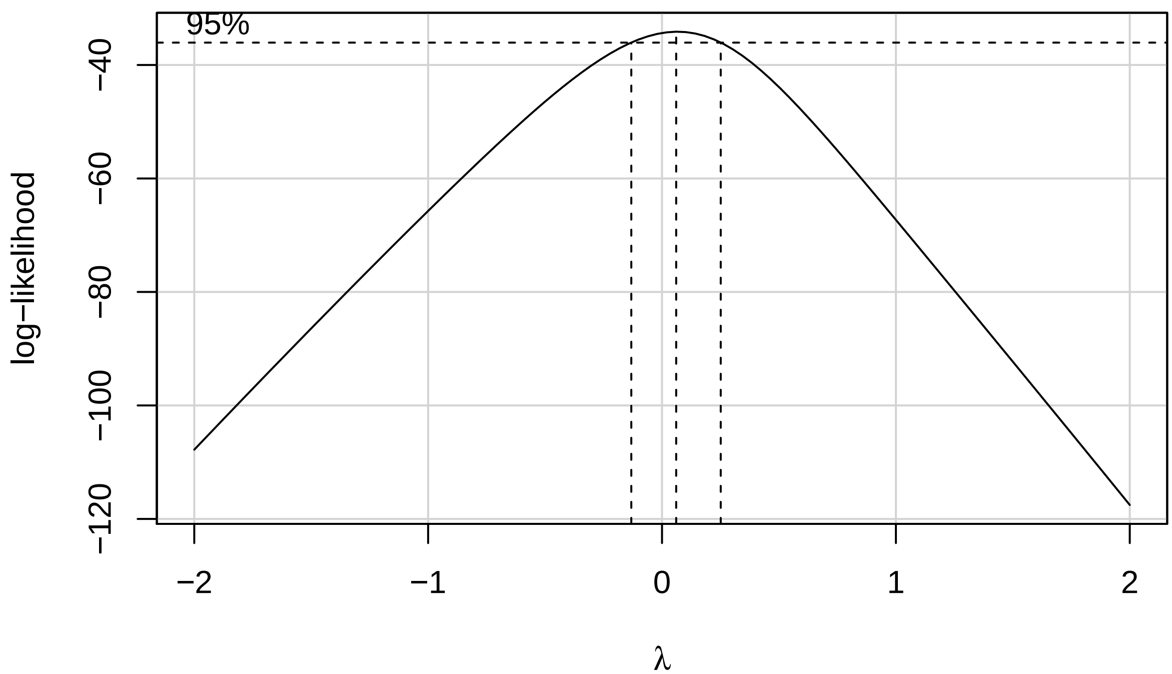
```
## [1] 3.084061
```

Since our test statistic is greater than our critical value (2.01) we determine that this point is influential.

Question 3

Part(a)

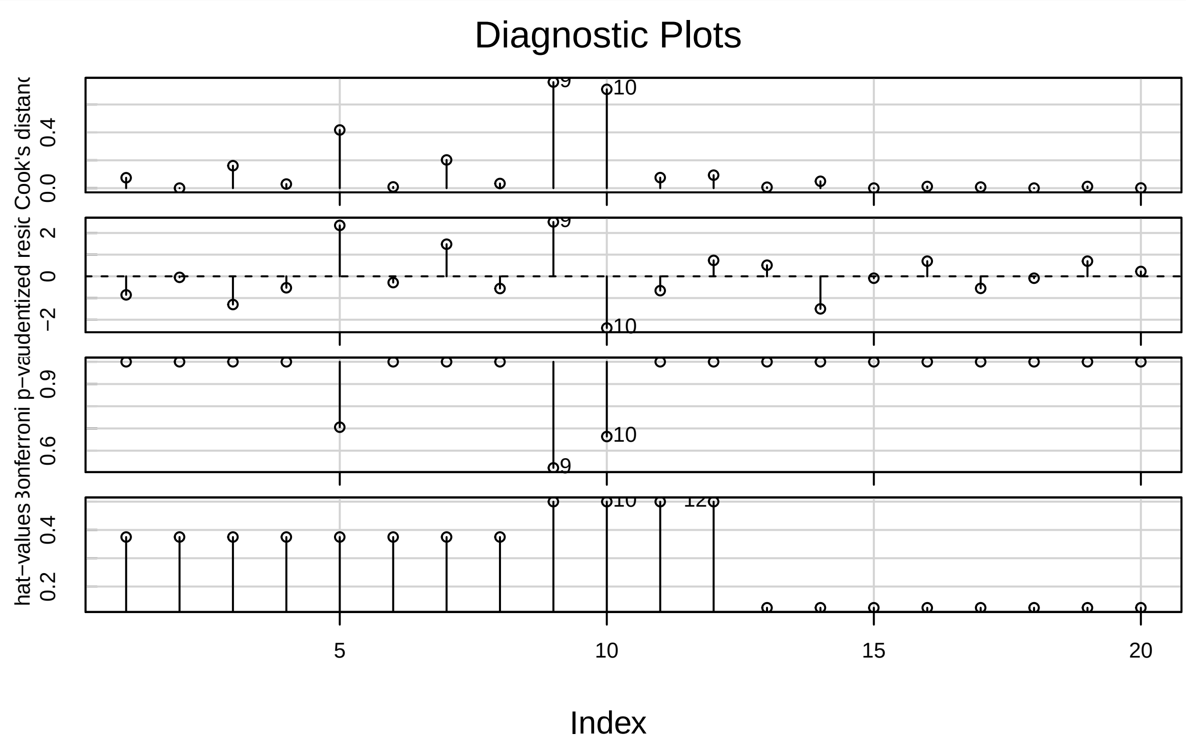
```r
data(lathe1)
Speed <- lathe1$Speed
Feed <- lathe1$Feed
Life <- lathe1$Life
fit1 <- lm(Life ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed*Feed)
boxCox(fit1)
```



Since 0 is within the interval, we know to log transform our response.

Part(b)

```r
fit2 <- lm(log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed*Feed)
influenceIndexPlot(fit2)
```



Here we see points 9 and 10 have the highest cooks distance, therefore we will remove them.

```r
new.lathe1 <- lathe1[-c(9, 10), ]
new.Speed <- new.lathe1$Speed
new.Feed <- new.lathe1$Feed
new.Life <- new.lathe1$Life
fit3 <- lm(log(new.Life) ~ new.Speed + new.Feed + I(new.Speed^2) + I(new.Feed^2) + new.Speed*new.Feed)
summary(fit2)
```

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

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = log(new.Life) ~ new.Speed + new.Feed + I(new.Speed^2) +
##     I(new.Feed^2) + new.Speed * new.Feed)
##
## 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 ***
## new.Speed          -1.43300    0.08241 -17.388 7.10e-10 ***
## new.Feed           -0.79023    0.06729 -11.743 6.15e-08 ***
## I(new.Speed^2)      0.28022    0.12363   2.267 0.042700 *
## I(new.Feed^2)       0.42244    0.09217   4.583 0.000629 ***
## new.Speed:new.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
```

Based off of our summary function we see that before removing the influencial points, our R squared value is 0.9702, and after we get an R squared of 0.9759. Therefore we can see the non-influential data has a slightly better fit than with the influential points.