# Homework 4 - STATS 513

Saad Ahmad

## Question 1

### Consider the stackloss data, with stack.loss being the response and the other three variables being predictors. Fit models with the following 3 methods respectively and Compare the results. In each case, comment on the significance of predictors.

```
library(faraway)
library(quantreg)
library(MASS)
```

## 1. Ordinary least squares

```
linearModel1 = lm(stack.loss ~ ., data = stackloss)
summary(linearModel1)
```

```
Call:
lm(formula = stack.loss ~ ., data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2377 -1.7117 -0.4551  2.3614  5.6978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -39.9197    11.8960  -3.356  0.00375 ** 
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 ** 
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,	Adjusted R-squared:  0.8983 
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

Acid.Conc. is not significant whereas the other variables such as Air.Flow & Water.Temp are.

## 2. Least absolute deviations

```
linearModel2 = rq(stack.loss ~ ., data = stackloss)
summary(linearModel2)
```

```
Call: rq(formula = stack.loss ~ ., data = stackloss)

tau: [1] 0.5

Coefficients:
            coefficients lower bd  upper bd
(Intercept) -39.68986    -41.61973 -29.67754
Air.Flow      0.83188      0.51278   1.14117
Water.Temp    0.57391      0.32182   1.41090
Acid.Conc.   -0.06087     -0.21348  -0.02891
>
```

All the co-efficients are now significant at 0.05 as all 95% confidence intervals doesn't contain 0.
The coefficients of Water.Temp and Acid.Conc. are changed significantly.

## 3. Huber's robust regression

```
linearModel3 = rlm(stack.loss ~ ., data = stackloss)
summary(linearModel3)
```

```
Call: rlm(formula = stack.loss ~ ., data = stackloss)
Residuals:
     Min       1Q   Median       3Q      Max
-8.91753 -1.73127  0.06187  1.54306  6.50163

Coefficients:
            Value    Std. Error t value
(Intercept) -41.0265   9.8073    -4.1832
Air.Flow      0.8294   0.1112     7.4597
Water.Temp    0.9261   0.3034     3.0524
Acid.Conc.   -0.1278   0.1289    -0.9922

Residual standard error: 2.441 on 17 degrees of freedom
```

The coefficients besides the Acid.Conc. are significant.

## Use diagnostic methods to detect any outliers or influential points. Remove these points and then use least squares. Compare the results.

Determining the outliers or influential points for the least squares method.

For outliers we will use the standard Bonferonni method as Professor Bo Wei mentioned in class. The p-value of the largest studentized residual is 0.00396 and the 0.05/17 quantile is 0.00294 which is not significant and we don’t have any outliers.

```
?rstudent

temp = rstudent(linearModel1)

### Compute p-value

2*(1-pt(max(abs(temp)), df = 17))

### compare to alpha/n

0.05/17
```

```
> temp = rstudent(linearModel1)
> ### Compute p-value
> 2*(1-pt(max(abs(temp)), df = 17))
[1] 0.003960491
> ### compare to alpha/n
> 0.05/17
[1] 0.002941176
```

For the influential points, we will use the Cook’s distance as discussed by Professor Bo Wei in the class. Taking out the least squares. The value of the coefficient on Water.Temp is significantly less than we had previously in the least square model. The coefficients on Air.Flow and Water.Temp are still significant.

```
cookDistance = cooks.distance(linearModel1)

stackloss_model1 = stackloss[-21,]

linearModel1_modified = lm(stack.loss ~ ., data = stackloss_model1)
```

```
summary(linearModel1)
```

```
Call:
lm(formula = stack.loss ~ ., data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2377 -1.7117 -0.4551  2.3614  5.6978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -39.9197    11.8960  -3.356  0.00375 ** 
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 ** 
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,	Adjusted R-squared:  0.8983 
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

---

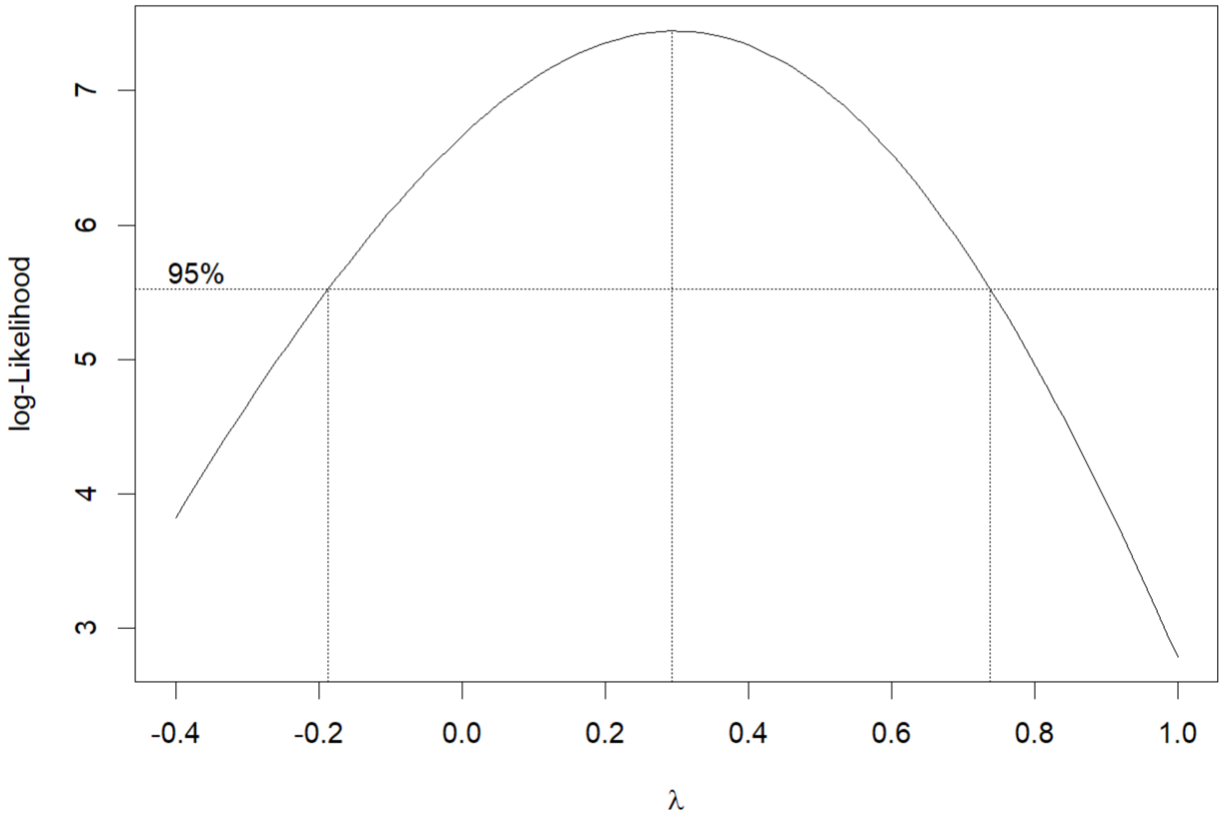
## Question 2

## Utilize the Box-Cox model to determine if there should be a transformation on the stack.loss variable in the stackloss data. If there should be one, read the estimate of λfrom the plot and fit a model with transformation

```
linearModel_BoxCox = boxcox(lm(stack.loss~.,data = stackloss), lambda
= seq(-.4, 1, by = .05))

linearModel_powlaw = lm(stack.loss^.33~. , data = stackloss)

summary(linearModel_powlaw)
```




```
Call:
lm(formula = stack.loss^0.33 ~ ., data = stackloss)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27371 -0.08081 -0.02840  0.13834  0.29781 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.4601505  0.5065532  -0.908  0.37637    
Air.Flow     0.0300941  0.0057425   5.241 6.64e-05 ***
Water.Temp   0.0564286  0.0156711   3.601  0.00221 ** 
Acid.Conc.  -0.0006291  0.0066553  -0.095  0.92579    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1381 on 17 degrees of freedom
Multiple R-squared:  0.9188,	Adjusted R-squared:  0.9045 
F-statistic: 64.14 on 3 and 17 DF,  p-value: 1.777e-09
```

This transformed model leaves us with the coefficient for the predictor Acid.Conc. as the only predictor with non-significant p-value