Stat500(Section002): Homework #7

Due on Nov.17, 2021 at 3:00pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

Part a

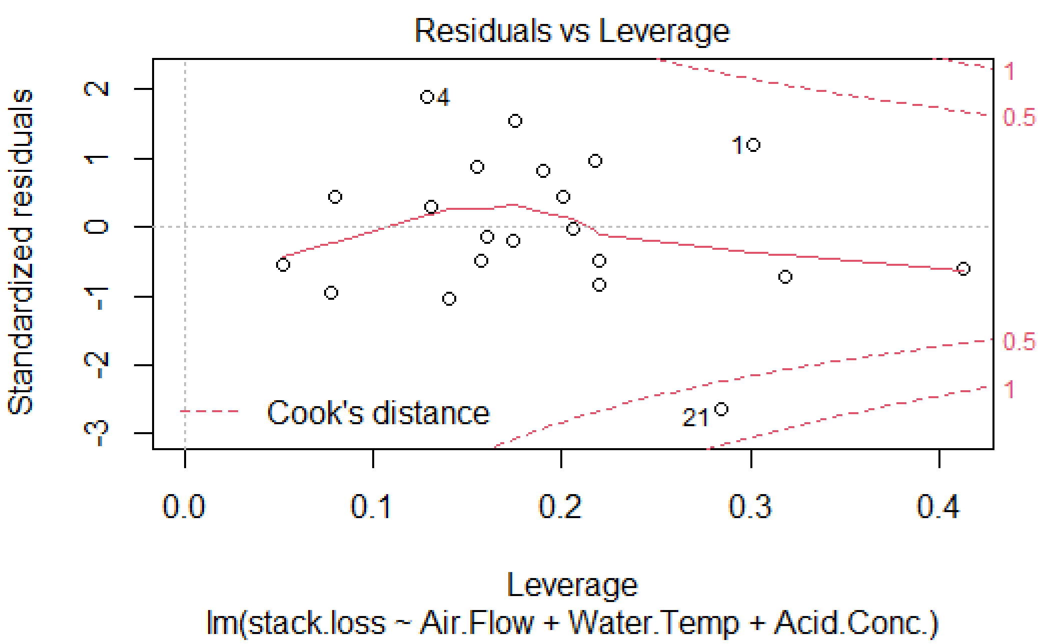
Using the following code to fit the different model

```
library(faraway)
library(quantreg)
library(MASS)
data("stackloss")
lsmod = lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
ladmod = rq(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
hubmod = rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
ltsmod = ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,
data=stackloss,nsamp="exact")
```

We will use the OLS model to check outlier and influential points only. Use the following code to check the outlier:

```
ti=rstudent(lsmod)
max(abs(ti))
which(abs(ti)==max(abs(ti)))
2*(1-pt(max(abs(ti)),df=21-3-1))
0.05/21
```

We get that the p-value is 0.00396 which is larger than $\frac{0.05}{21} = 0.00238$. Hence We don't think model has outlier. However, For the influential points, we have the result



From the figure we can see that 21 is the case number of influential points. Hence, we need to remove number 21. Using the folloing code:

```
lsmod_rm = lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,
data=stackloss,subset = -c(21))
```

```
ladmod_rm = rq(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,
data=stackloss,subset= -c(21))
hubmod_rm = rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,
data=stackloss,subset= -c(21))
ltsmod_rm = ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,
data=stackloss,nsamp="exact",subset= -c(21))
```

Now let us compare two models one by one.

(a) Least squares

```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    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
```

```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss, subset = -c(21))

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0449 -2.0578  0.1025  1.0709  6.3017 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -43.7040     9.4916  -4.605 0.000293 ***
Air.Flow      0.8891     0.1188   7.481 1.31e-06 ***
Water.Temp    0.8166     0.3250   2.512 0.023088 *  
Acid.Conc.   -0.1071     0.1245  -0.860 0.402338    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.569 on 16 degrees of freedom
Multiple R-squared:  0.9488,    Adjusted R-squared:  0.9392 
F-statistic: 98.82 on 3 and 16 DF,  p-value: 1.541e-10
```

We can see that, after we remove the influential point, the R^2 and $Adj\ R^2$ gets higher. And all the coefficient in the model change a lot. Especially for Water.Temp. Changed about 30 percent.

(b) Least absolute deviations

```
Call: rq(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    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
```

```
Call: rq(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss, subset = -c(21))

tau: [1] 0.5

Coefficients:
            coefficients lower bd  upper bd
(Intercept) -39.98645    -54.13730 -30.21338
Air.Flow      0.83469      0.82512   1.16654
Water.Temp    0.56369      0.26671   1.10718
Acid.Conc.   -0.05691     -0.39795  -0.03492
```

We can see that all the coefficient in the model only change a little bit. However, the upper and lower bound of estimated value changes a lot.

(c) Huber method

```
Call: rlm(formula = stack.loss ~ Air.Flow + Water.Temp +
Acid.Conc.,
    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
```

```
Call: rlm(formula = stack.loss ~ Air.Flow + Water.Temp +
Acid.Conc.,
    data = stackloss, subset = -c(21))
Residuals:
    Min      1Q  Median      3Q     Max
-2.9245 -1.5940  0.1337  1.0254  6.8283

Coefficients:
            Value    Std. Error t value
(Intercept) -42.8415   8.6193    -4.9704
Air.Flow      0.9184   0.1079     8.5093
Water.Temp    0.6854   0.2952     2.3222
Acid.Conc.   -0.1078   0.1131    -0.9529

Residual standard error: 2.273 on 16 degrees of freedom
```

In this method, Water.Temp again become the variable with most change. And all the coefficient changes. However, it changes more than Least absolute deviations less than Least squares.
(d) Least trimmed squares

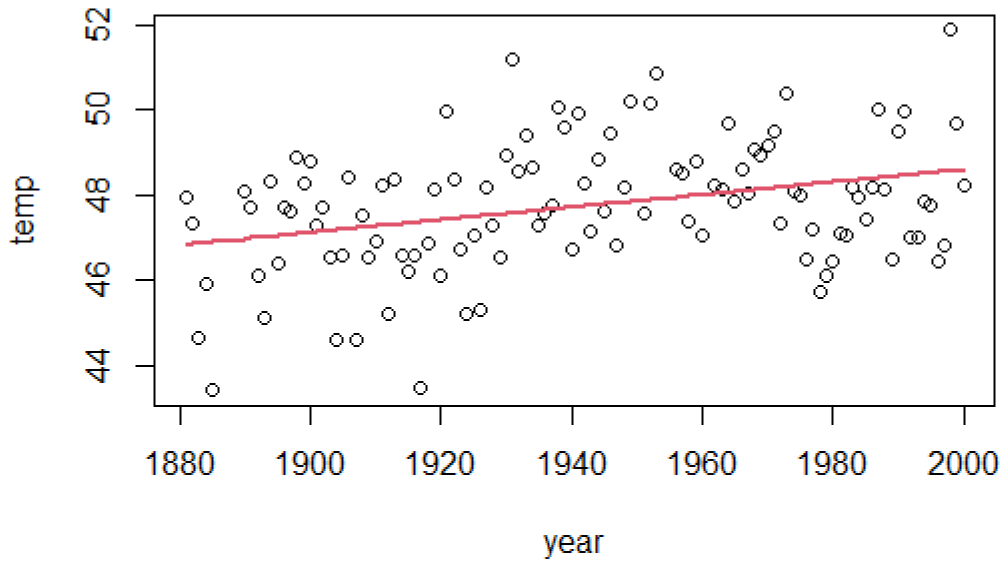
```
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc.
-3.580556e+01  7.500000e-01  3.333333e-01  3.489094e-17
```

```
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc.
-3.580556e+01  7.500000e-01  3.333333e-01  3.489094e-17
```

We can find that for Least trimmed squares, the estimated values of coefficient do not change at all.

Problem 2

part a
We plot the figure with year and temp, The red line shows that fitted value from model temp as response and year as predictor.



The plot shows some kind of linear trends of temp and year. Hence we think thereis an increasing linear trend in temperature.

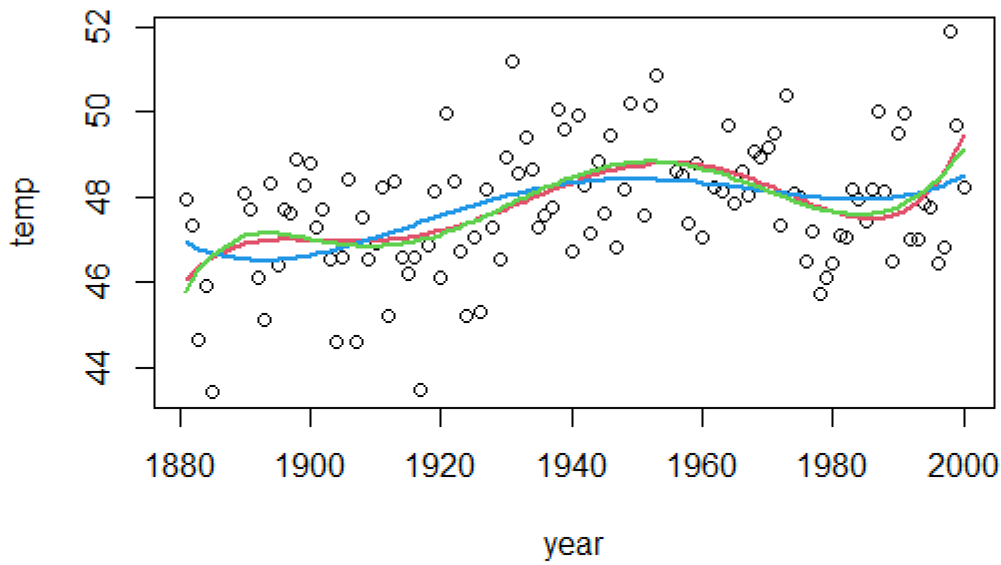
part b

We get the result of "interval(g)"

```
Approximate 95% confidence intervals

 Coefficients:
                      lower       est.       upper
(Intercept)      47.4287871 47.7300649 48.03134268
poly(year, 5)1    2.1415832  5.3237087  8.50583423
poly(year, 5)2   -6.3928177 -3.2338068 -0.07479587
poly(year, 5)3   -3.7276583 -0.5717096  2.58423918
poly(year, 5)4   -0.5027538  2.6575859  5.81792564
poly(year, 5)5    0.1951763  3.3389011  6.48262577
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
           lower      est.     upper
Phi1 -0.06234504 0.1476418 0.3451115
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
   lower     est.    upper
1.211960 1.395560 1.606974
```

We can see that the estimated value of Phi is 0.1476, And the the 95% lower bound of Phi is even less than 0, the 95% upper bound of Phi is only 0.3451. Therefore, we can say that we should use of LS without considering correlations among errors.

Part c



In this figure, the blue line shows the fitted value of 4-th degree model, red line shows the fitted value of 5-th degree model,green line shows the fitted value of 6-th degree model. We can see that, red line and green line are nearly same. Hence using 6-degree model shows no better than 5-degree model. Also, let us see the summary of 5-th degree and 6-th degree model.

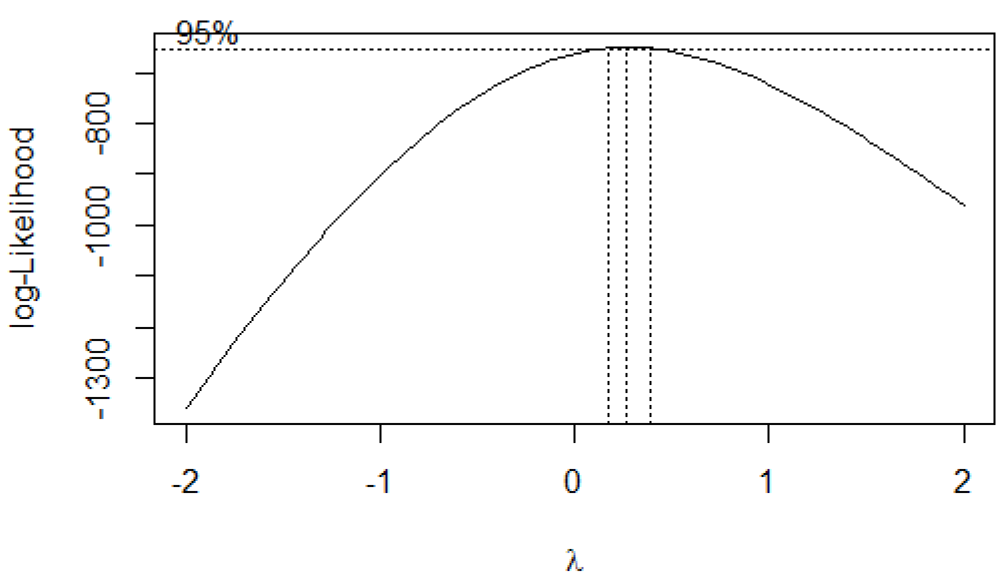
```
Call:
lm(formula = temp ~ poly(year, 5))

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6176 -0.8192 -0.1745  1.0038  3.3797 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     47.7315     0.1306 365.370  < 2e-16 ***
poly(year, 5)1   5.3613     1.3826   3.878 0.000183 ***
poly(year, 5)2  -3.1919     1.3826  -2.309 0.022899 *  
poly(year, 5)3  -0.4907     1.3826  -0.355 0.723327    
poly(year, 5)4   2.6345     1.3826   1.906 0.059415 .  
poly(year, 5)5   3.5041     1.3826   2.535 0.012721 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.383 on 106 degrees of freedom
Multiple R-squared:  0.2237,	Adjusted R-squared:  0.1871 
F-statistic:  6.11 on 5 and 106 DF,  p-value: 5.183e-05
```

```
Call:
lm(formula = temp ~ poly(year, 6))

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5009 -0.9005 -0.2233  1.0259  3.3189 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      47.7315     0.1308 364.910  < 2e-16 ***
poly(year, 6)1    5.3613     1.3843   3.873 0.000187 ***
poly(year, 6)2   -3.1919     1.3843  -2.306 0.023086 *  
poly(year, 6)3   -0.4907     1.3843  -0.355 0.723668    
poly(year, 6)4    2.6345     1.3843   1.903 0.059757 .  
poly(year, 6)5    3.5041     1.3843   2.531 0.012845 *  
poly(year, 6)6   -1.1853     1.3843  -0.856 0.393817    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.384 on 105 degrees of freedom
Multiple R-squared:  0.2291,	Adjusted R-squared:  0.1851 
F-statistic: 5.201 on 6 and 105 DF,  p-value: 0.0001013
```

We can see that they have nearly same $AdjR^2$. And 6-th degree model even has a lower $AdjR^2$.Incresing model degree from 5 to 6 has no benefit. And we can see that the x^6 here is not sinificant. There is no reason to use 6-th degree model.
Hence we get the conclusion that there is no need to use an polynomial with an even higher order to model this dataset.

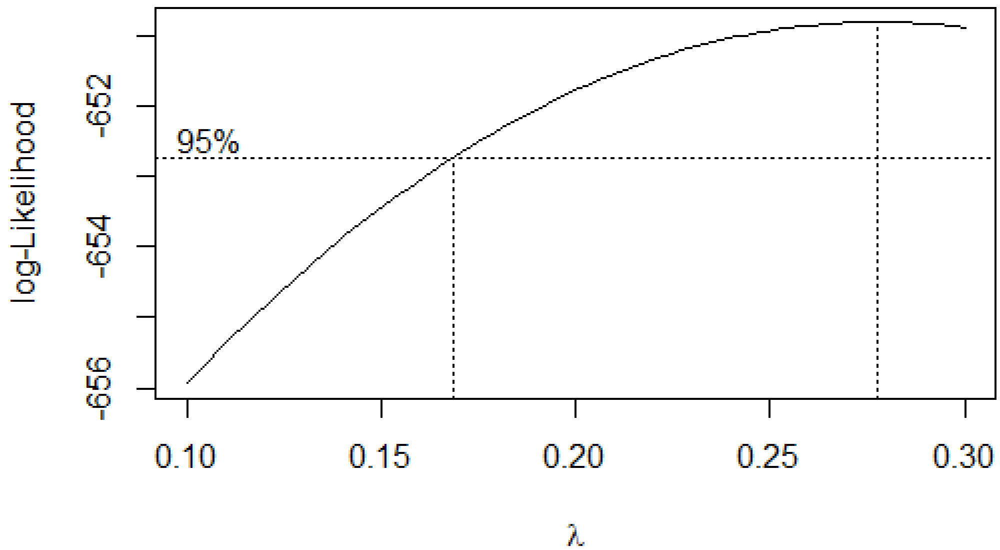
Problem 3

First, we show the overall plot of boxcox method to see a detail range of best number.



Problem 3 continued on next page. . .

From the plot, we can see that the best transform number stay around 0.1 to 0.35. And we should use transfomation of course. Let us plot a more detailed figure.



Then we test all the number between 0.275 to 0.280 with step 0.001 to get the result that when $\lambda = 0.279$, we get the max log-Likelihood. Hence we get the transformation:

$$g_\lambda(y) = \frac{y^{0.279} - 1}{0.279}$$