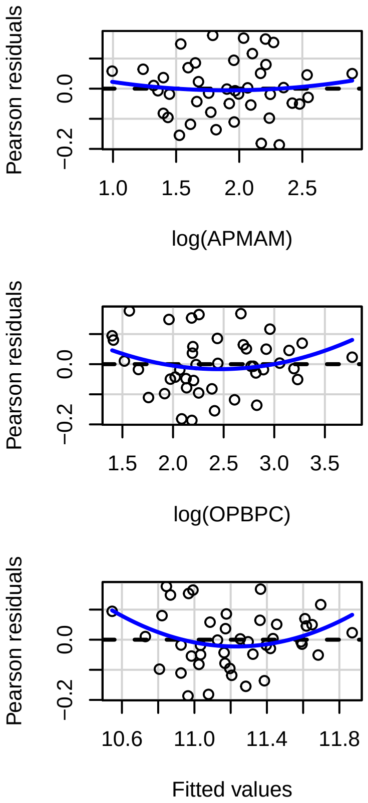
S631 HW9

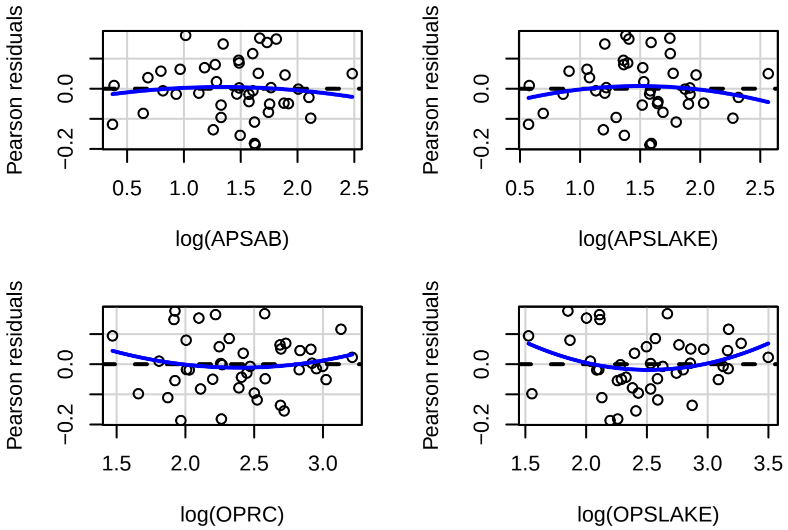
Shibi He

2. ALR 9.8

```r
library(alr4)
m1 = lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
            log(OPBPC) + log(OPRC) + log(OPSLAKE), data= water)

rp=residualPlots(m1)
```





```
##                Test stat Pr(>|Test stat|)
## log(APMAM)        0.4499          0.65553
## log(APSAB)       -0.4647          0.64502
## log(APSLAKE)     -0.8525          0.39976
## log(OPBPC)        1.3848          0.17487
## log(OPRC)         0.8387          0.40735
## log(OPSLAKE)      1.6295          0.11217
## Tukey test        1.8386          0.06597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

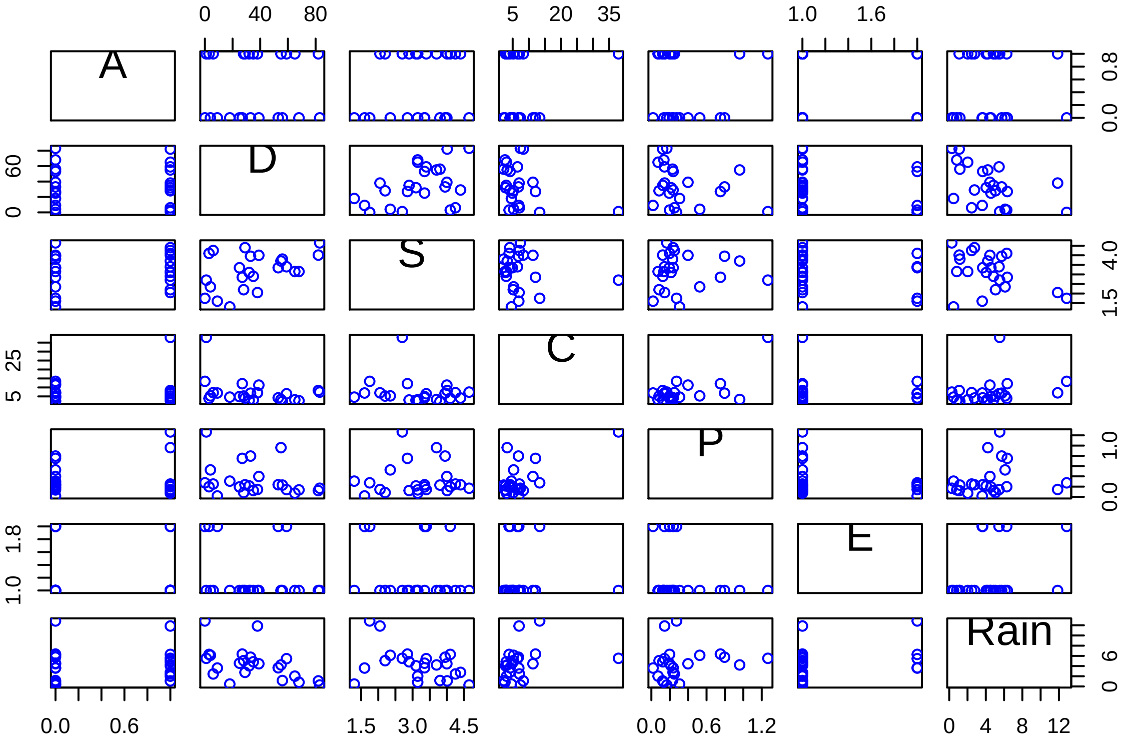
All of the plots versus individual regressors look like null plot and no visual evidence of curvature. The last plot of residuals versus the fitted values only shows a hint of curvature. But the Tuckey test gives a pvalue of 0.066 (>0.05), so we cannot reject the original mean function. Actually, all of the tests have pvalues greater than 0.05, providing no evidence against the mean function.

```
# Alternatively,
# yhat=fitted(m1)
# summary(lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
#log(OPBPC) + log(OPRC) + log(OPSLAKE)+I(yhat^2), water))$coefficients
#
# # pvalue using standard normal distribution
# 2*(1-pnorm(1.8386))
# p= 0.06597404
```

3. ALR 9.18

i. Draw scatter plots

```
scatterplotMatrix(~A+D+S+C+P+E+Rain, data=cloud, smooth = F,
                  regLine = F, diagonal = F)
```



```
View(cloud$C)
```

The scatter plots show that the relationship between *Rain* and *S*, *C*, *P* seems to be somewhat curved, so transformations of these predictors might be useful. The relationship between predictors themselves does not have any specific pattern.

ii. Transformation for the predictors

```
bc1 = powerTransform(cbind(S, C, P) ~ 1, cloud)
summary(bc1)
```

```
## bcPower Transformations to Multinormality
##   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## S    1.2888           1       0.0015       2.5760
## C   -0.3939           0      -0.9665       0.1787
## P    0.1297           0      -0.2070       0.4664
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df     pval
## LR test, lambda = (0 0 0) 7.306579  3 0.062742
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1) 47.30214  3 2.9975e-10
```

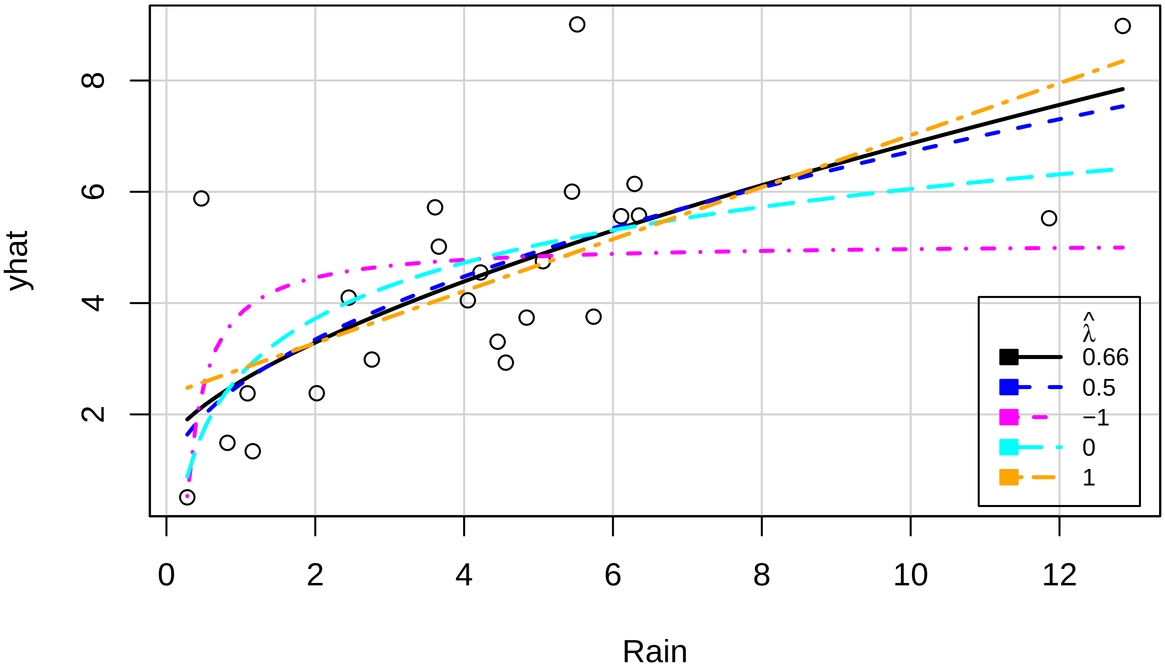
```
testTransform(bc1, c(1, 0, 0))
```

```
##                                  LRT df   pval
## LR test, lambda = (1 0 0) 3.063775  3 0.3819
```

Using Box-Cox method, the suggested transformation is no transformation for S, and log transformation for C and P. The likelihood ratio test for $\lambda = (1\ 0\ 0)$ gives a pvalue of 0.38(>0.05), providing no evidence against these transformations.

iii. Transformation for the response

```
m1 = lm(Rain ~ A + D + S + log(C) + log(P) + E, data=cloud)
p1=inverseResponsePlot(m1, c(0.5, -1, 0, 1))
```



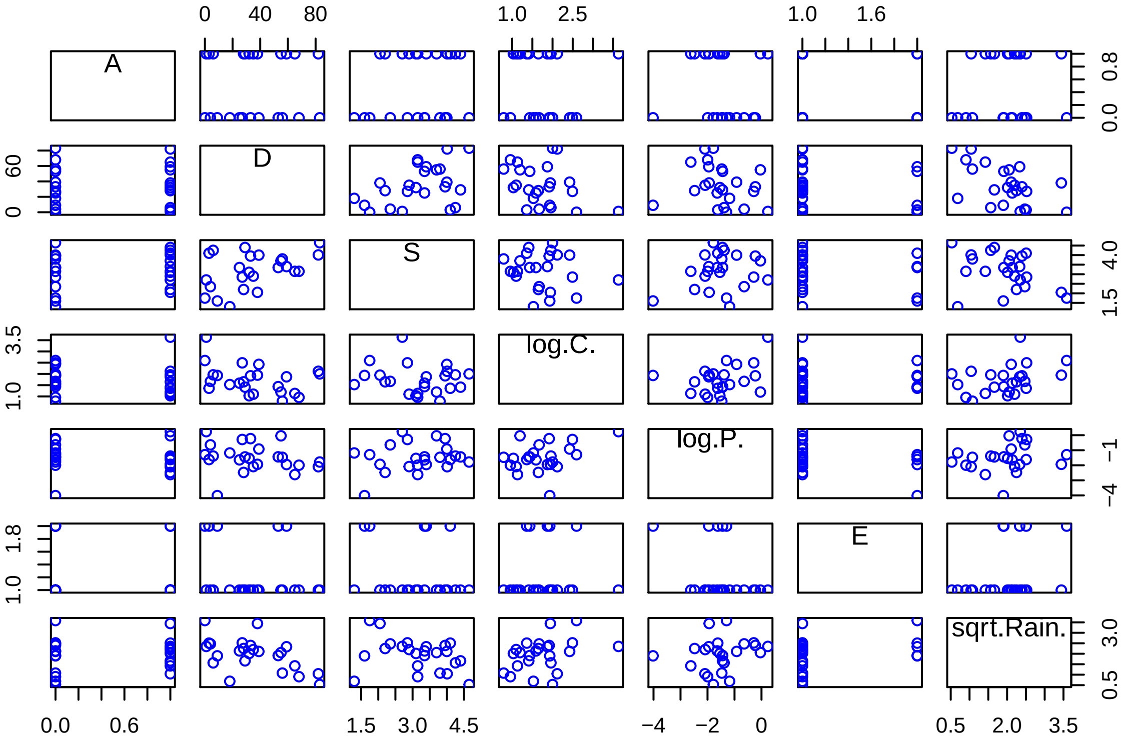
```
p1
```

```
##      lambda      RSS
## 1 0.6566795 53.67112
## 2 0.5000000 54.07896
## 3 -1.0000000 80.19895
## 4  0.0000000 60.64466
## 5  1.0000000 55.34328
```

The fitted line with $\lambda = 0.5$ is very close to the best fitting line($\hat{\lambda} = 0.66$), and the RSS(λ=0.5) is also very small, suggesting taking square root of *Rain* might be reasonable.

iv. Draw scatter plots again to check if transformation makes improvment:

```
scatterplotMatrix(~A + D + S + log(C) + log(P) + E + sqrt(Rain),
                  data=cloud, smooth = F, regLine = F, diagonal = F)
```



The relationship between response (*Rain*) and the predictors now become relatively more close to linear.

v. Fit the model

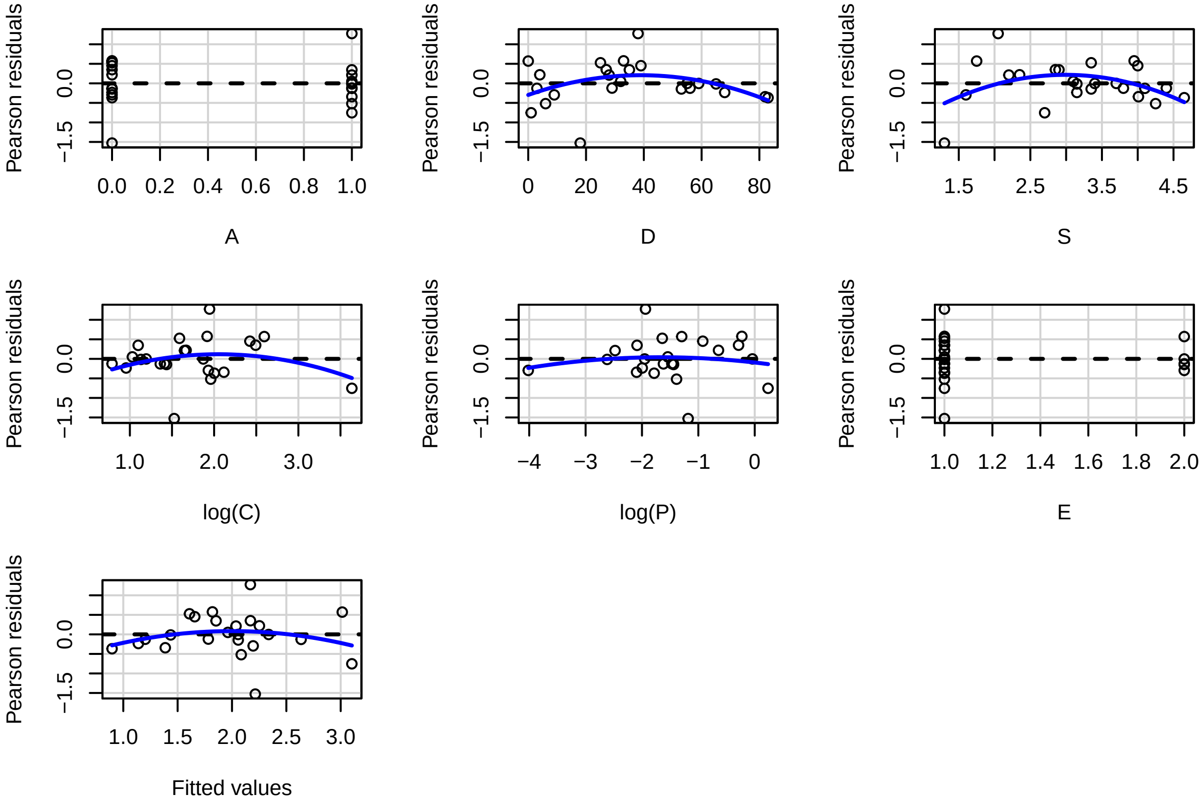
```
m2 = lm(sqrt(Rain) ~ A + D + S + log(C) + log(P) + E, data=cloud)
summary(m2)
```

```
##
## Call:
## lm(formula = sqrt(Rain) ~ A + D + S + log(C) + log(P) + E, data = cloud)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52860 -0.24875 -0.00868  0.34746  1.27457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.085128   0.881154   2.366   0.0301 *
## A            0.406592   0.268678   1.513   0.1486
## D           -0.007798   0.006611  -1.180   0.2544
## S           -0.216114   0.175098  -1.234   0.2339
## log(C)       0.138919   0.236496   0.587   0.5647
## log(P)       0.253309   0.182668   1.387   0.1834
## E            0.637495   0.360000   1.771   0.0945 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6295 on 17 degrees of freedom
## Multiple R-squared:  0.4953, Adjusted R-squared:  0.3172
## F-statistic: 2.781 on 6 and 17 DF,  p-value: 0.04519
```

The results show that the p-value for all predictors are greater than 0.05, suggesting none of the predictors has statistically significant effects on rainfall. Next, I turn to model checking.

vi. Residual plots and test for curvature

```
rp = residualPlots(m2)
```



```
##              Test stat Pr(>|Test stat|)
## A              -2.0342          0.05886 .
## D              -2.0132          0.06123 .
## S              -2.0170          0.06079 .
## log(C)         -1.2685          0.22278
## log(P)         -0.5802          0.56985
## E              -0.3452          0.73441
## Tukey test   -1.0538          0.29195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual plots look like null plots and do not have any visual evidence for curvature. Moreover, the Tukey test gives a p-value of 0.29, providing no evidence against the mean function.

vii. Check for outliers

```r
outlierTest(m2, cutoff = 0.8)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 7  -4.074011         0.00088343     0.021202
## 15  2.674945         0.01660200     0.398440
```

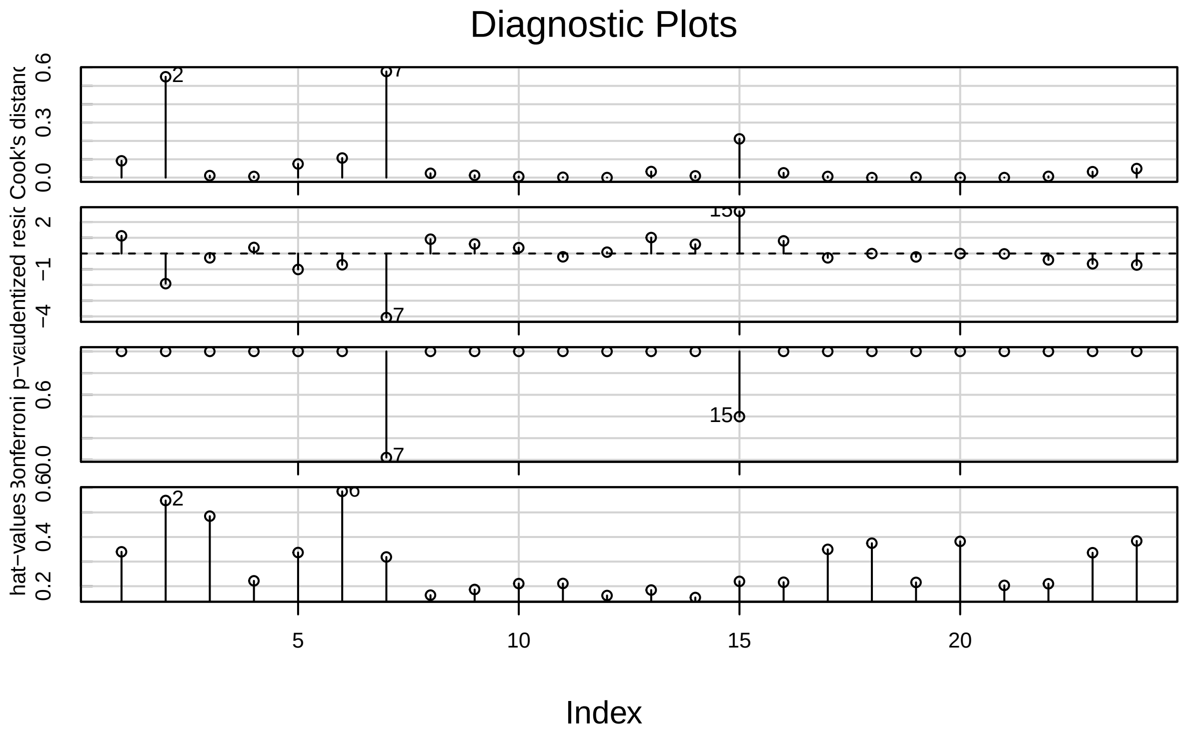
The 15th and 7th observation have the largest standardized residuals with Bonferroni p-values of 0.021 (<0.05) and 0.398 (>0.05), respectively. So observation 7 is an outlier while observation 15 is not an outlier.

viii. Check for influential cases:

```r
cooks.distance(m2)
```

```
##            1            2            3            4            5 
## 9.137290e-02 5.497754e-01 1.080272e-02 6.358372e-03 7.462682e-02 
##            6            7            8            9           10 
## 1.067414e-01 5.786154e-01 2.338554e-02 1.245924e-02 5.499525e-03 
##           11           12           13           14           15 
## 1.793234e-03 2.103648e-04 3.313347e-02 9.324030e-03 2.111816e-01 
##           16           17           18           19           20 
## 2.590467e-02 6.216649e-03 1.288411e-06 1.968548e-03 2.986909e-06 
##           21           22           23           24 
## 2.420979e-05 6.604364e-03 3.219634e-02 4.924715e-02
```

```r
influenceIndexPlot(m2)
```



Observation 2 and 7 have the largest Cook's ditance (0.55 and 0.58), so they are the relatively more influential observations. Next, I remove these two influential observations and fit the model again to see if they have large impact on the estimated coefficients.

```
cloud2 = cloud[c(-2, -7),]
m3 = lm(sqrt(Rain) ~ A + D + S + log(C) + log(P) + E, data=cloud2)
summary(m3)
```

```
##
## Call:
## lm(formula = sqrt(Rain) ~ A + D + S + log(C) + log(P) + E, data = cloud2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48129 -0.24934 -0.01312  0.18038  0.71677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.471124   0.572577   6.062 2.17e-05 ***
## A            0.494862   0.161760   3.059 0.007952 **
## D           -0.007378   0.003687  -2.001 0.063796 .
## S           -0.532603   0.110482  -4.821 0.000225 ***
## log(C)       0.299872   0.163119   1.838 0.085892 .
## log(P)       0.454706   0.107427   4.233 0.000724 ***
## E            0.430960   0.203338   2.119 0.051148 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3509 on 15 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.7759
## F-statistic: 13.12 on 6 and 15 DF,  p-value: 3.213e-05
```

After removing the influential observations, the coefficients of A, S, and $log(P)$ become significant, that is, cloud seeding is effctive in increasing the rainfall.

In summary, using all data, we do not find evidence that cloud seeding is effective in increasing rainfall. But this might be largely attributed to the two influential observations. Because when we remove the influential observations (day 2 and day 7), the effect of cloud seeding becomes positive and statistically significant.