Assignment 9

Chase Baggett

November 30, 2017

9.1

Power Transform

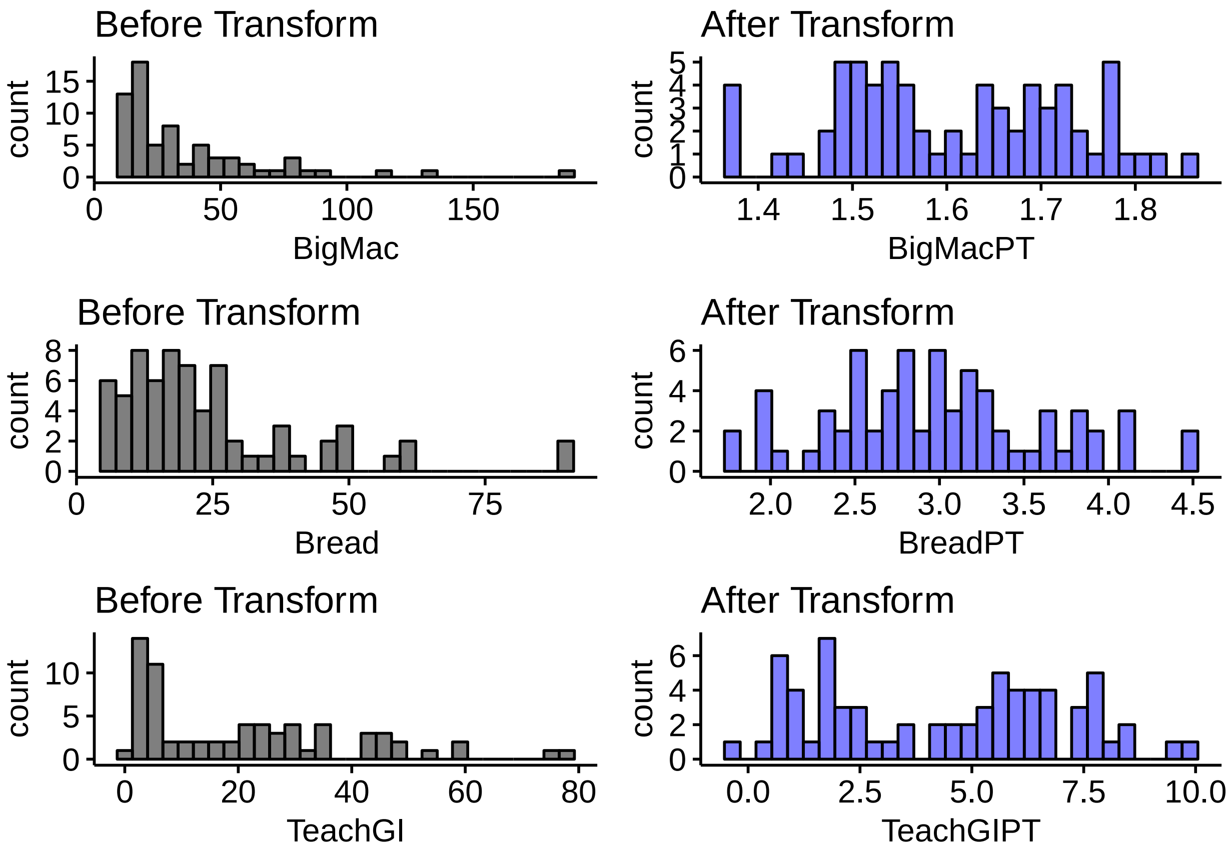
```
p_trans <- powerTransform(cbind(BigMac2003$BigMac,BigMac2003$Bread,BigMac2003$TeachGI) ~ 1)
summary(p_trans)$result
```

```
##     Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Y1 -0.4021272       -0.50   -0.6318245   -0.1724300
## Y2 -0.1619874        0.00   -0.4559858    0.1320110
## Y3  0.2692366        0.33    0.1104494    0.4280238
```

Round Values

```
pt <- function(x,lambda){if(lambda==0){log(x)}else{(x^lambda-1)/lambda}}
BigMac2003$BigMacPT <- (BigMac2003$BigMac^-.5 - 1)/-.5
#P189 says take log if zero.
BigMac2003$BreadPT <- log(BigMac2003$Bread)
BigMac2003$TeachGIPT <- (BigMac2003$TeachGI ^.33 -1)/.33

fit_original <- lm(BigMacPT ~ BreadPT + TeachGIPT,data=BigMac2003)
p1 <- gghistogram(BigMac2003,x="BigMac",fill="black") + ggtitle("Before Transform")
p2 <- gghistogram(BigMac2003,x="BigMacPT",fill="blue") + ggtitle("After Transform")
p3 <- gghistogram(BigMac2003,x="Bread",fill="black") + ggtitle("Before Transform")
p4 <- gghistogram(BigMac2003,x="BreadPT",fill="blue") + ggtitle("After Transform")
p5 <- gghistogram(BigMac2003,x="TeachGI",fill="black") + ggtitle("Before Transform")
p6 <- gghistogram(BigMac2003,x="TeachGIPT",fill="blue") + ggtitle("After Transform")
grid.arrange(p1,p2,p3,p4,p5,p6,ncol=2)
```



9.1.1

```
testTransform(p_trans,c(1,1,1))
```

```
##                                LRT df pval
## LR test, lambda = (1 1 1) 243.018  3    0
```

```
testTransform(p_trans,c(-1/3,0,1/3))
```

```
##                                        LRT df      pval
## LR test, lambda = (-0.33 0 0.33) 2.460323  3 0.4825043
```

9.1.2

The units for TeachGI are are income in US dollars, but for countries all over the world. The units for bread are minutes of labor to purchase 1kg of Bread. Both of these things are expected to be very non-normally distributed because of massive economic differences and the difference purchasing power parity in different countries.

9.2

We round it to -.5

```
fit <- lm(BigMac~Bread + TeachGI,data=BigMac2003)
fit_bc <- powerTransform(fit,family="bcPower")
fit_bc_summary <- summary(fit_bc)
fit_bc_summary$result[1,1]
```

```
## [1] -0.5141583
```

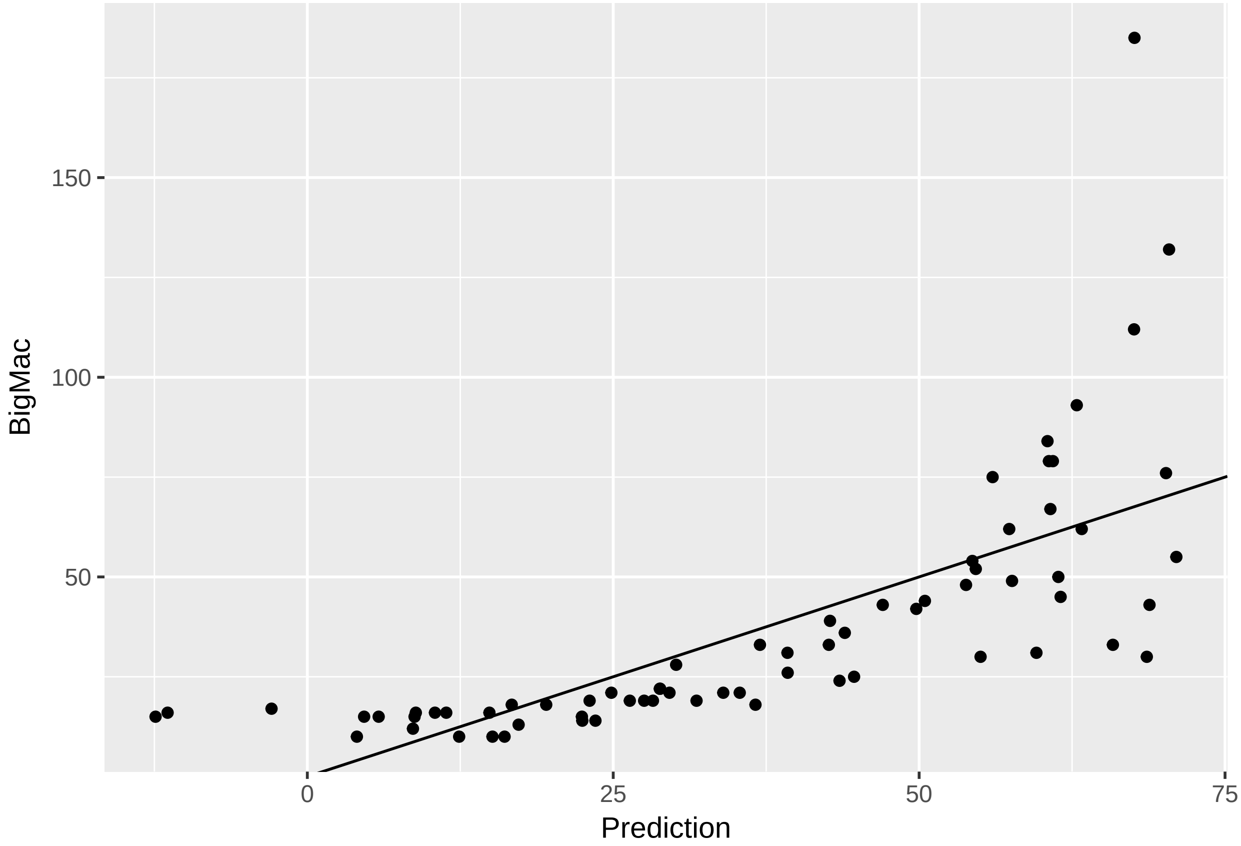
```
BigMac2003$BigMac92 <- pt(BigMac2003$BigMac,fit_bc_summary$result[1,1])
fit_bc_applied <- lm(BigMac92 ~ Bread + TeachGI,data=BigMac2003)
```

9.3

```
t_tog <-  summary(powerTransform(cbind(BigMac2003$Bread,BigMac2003$TeachGI) ~ 1))$result


#rounded power
BigMac2003$Bread_new <- pt(BigMac2003$Bread,t_tog[1,2])
BigMac2003$TeachGI_new <-  pt(BigMac2003$TeachGI,t_tog[2,2])

transformed_fit <- lm(BigMac ~ Bread_new + TeachGI_new,data=BigMac2003)
BigMac2003$Prediction <- predict(transformed_fit)
ggplot(BigMac2003,aes(x=Prediction,y=BigMac)) + geom_point() + geom_abline(slope=1,intercept = 0)
```

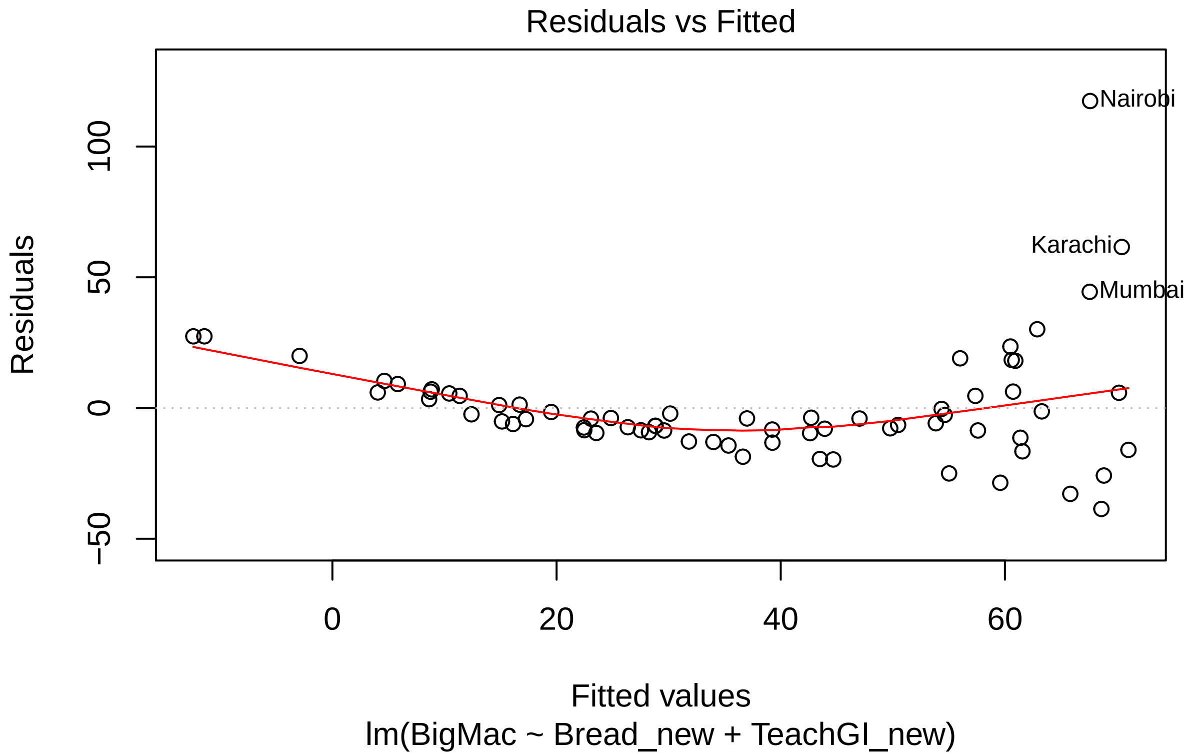


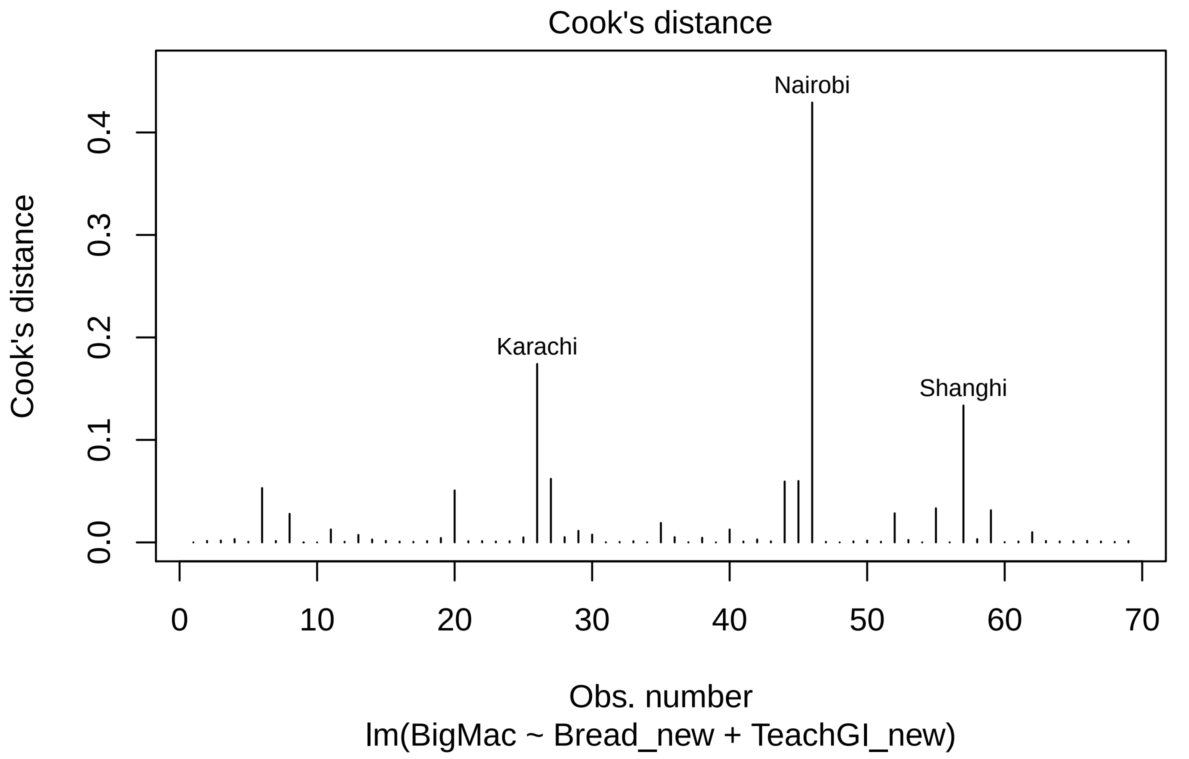
Transforming the predictors and not the response leaves us with alot of remaining non-constant variance.

9.4

Our Model from 9.3 has significant non-constant variance, but the relationship appears linear. We have three points with considerible influence.

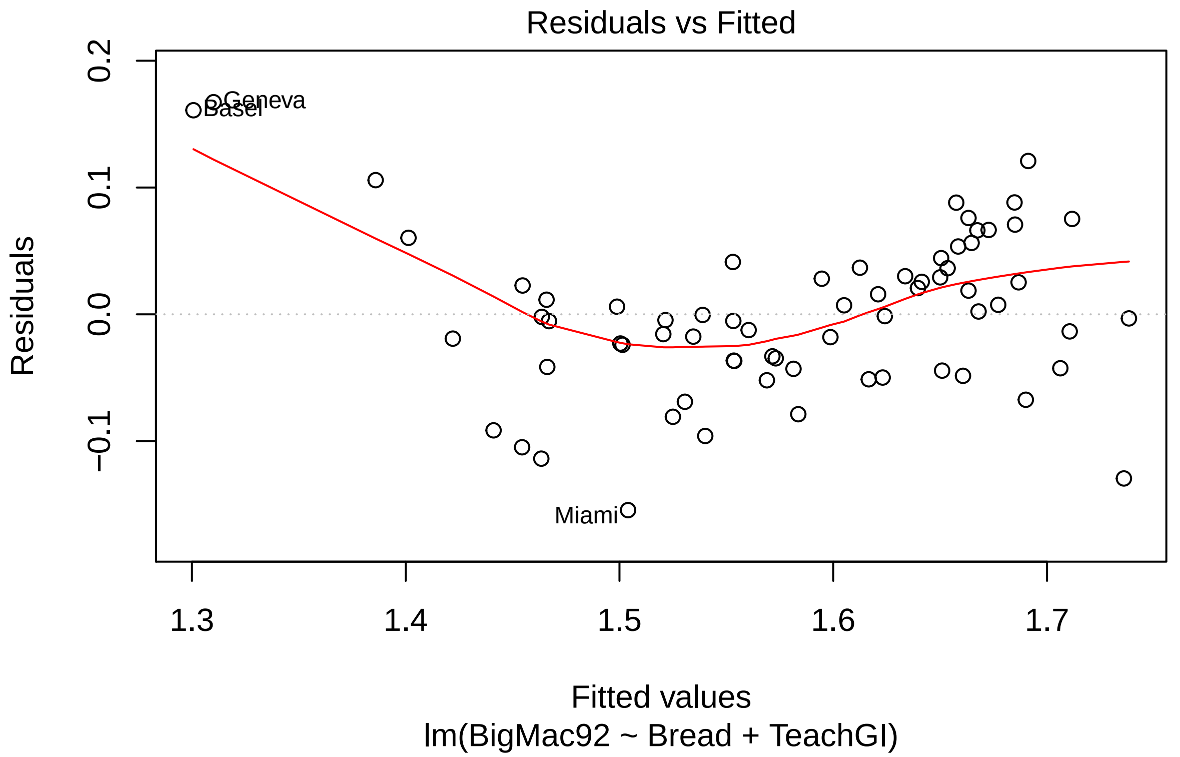
```r
plot(transformed_fit,which=c(1,4))
```

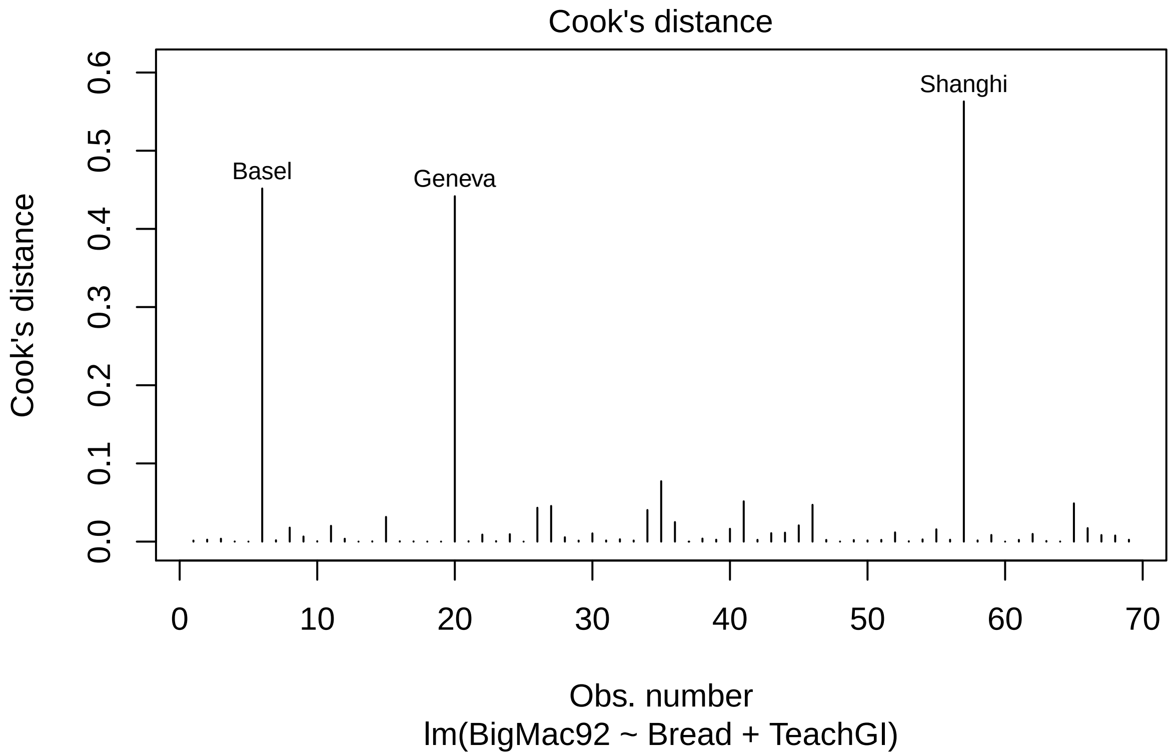

Residuals vs Fitted
Residuals
100
50
0
−50
0
20
40
60
Nairobi
Karachi
Mumbai
Fitted values
lm(BigMac ~ Bread_new + TeachGI_new)


Cook's distance
Cook's distance
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
50
60
70
Nairobi
Karachi
Shanghi
Obs. number
lm(BigMac ~ Bread_new + TeachGI_new)

Our model from 9.2 has solved this by power transforming the response, but the plot suggests our relationship is not linear. We have three points with significant influence, only one of which is the same. The influence is consideribly higher.

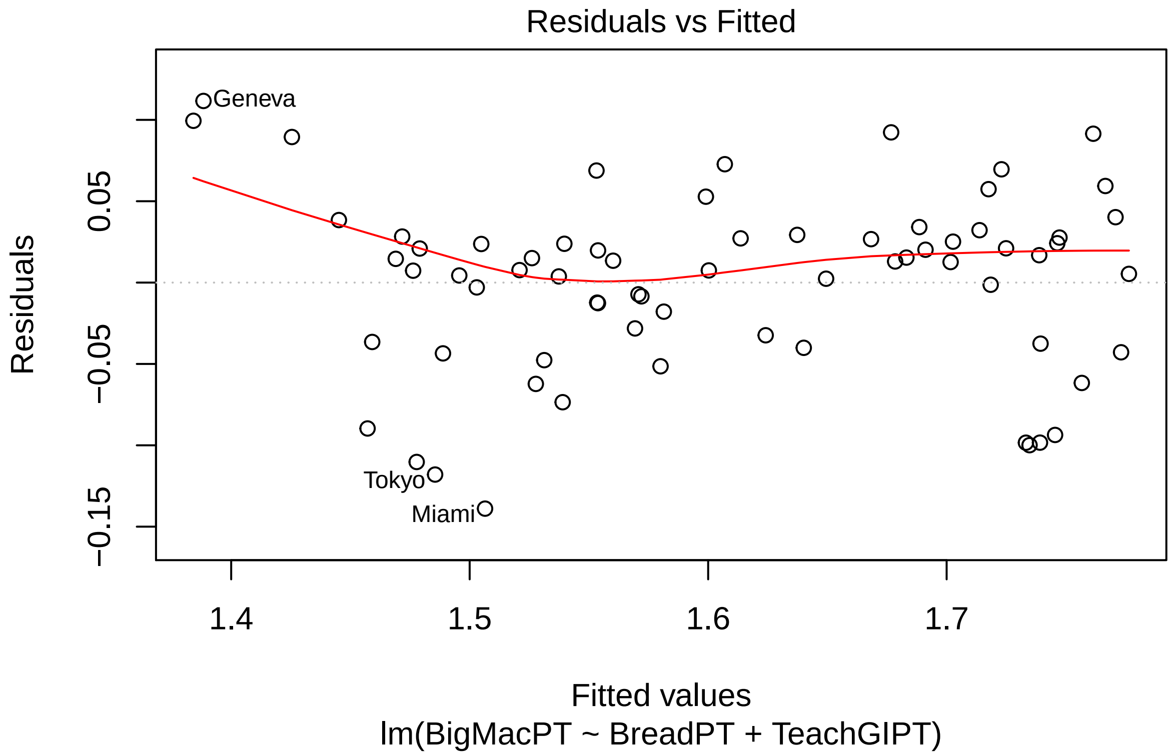
```r
plot(fit_bc_applied,which=c(1,4))
```

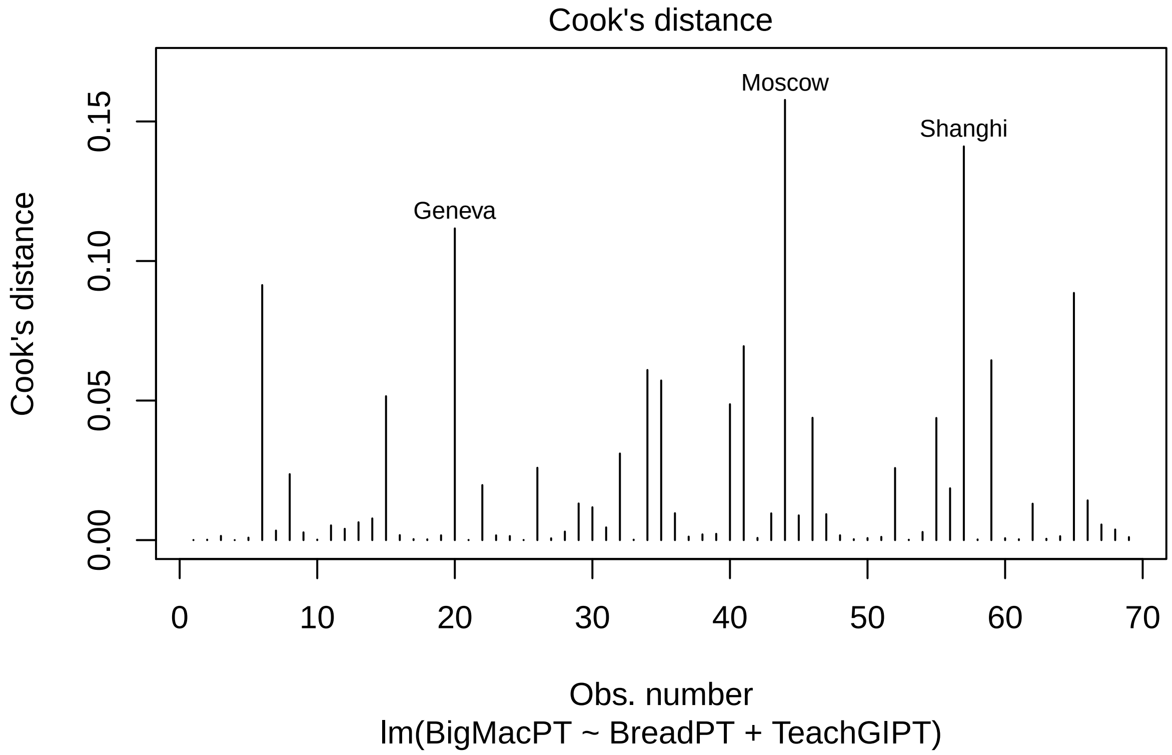




In 9.1, we still had some nonlinearity, but it is the most linear with the most constant variance. I would choose this model. There are alot more points with higher influence, but that may not be a bad thing– we may just be learning alot from those points.

```r
plot(fit_original,which=c(1,4))
```


Residuals vs Fitted
Residuals
0.05
−0.05
−0.15
1.4
1.5
1.6
1.7
Geneva
Tokyo
Miami
Fitted values
lm(BigMacPT ~ BreadPT + TeachGIPT)



9.5

We use the Bonferonni P. R returns NA because the P is over 1. But it would hypothetically be the unadjusted P * the # of tests performed, so I've manually produced the illogical result of a P value over 1. It is not an outlier because we generated it from the data, but might be had we picked it without looking at the data.

```
BigMac2003$BigMac95 <- pt(BigMac2003$BigMac,-1/3)
BigMac2003$Bread95 <- pt(BigMac2003$Bread,0)
BigMac2003$TeachGI95 <- pt(BigMac2003$TeachGI,1/3)

hyp_fit <- lm(BigMac95 ~ Bread95 + TeachGI95,data=BigMac2003)
outlierTest(hyp_fit)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##        rstudent unadjusted p-value Bonferonni p
## Miami -2.283701           0.025665           NA
```

```
outlierTest(hyp_fit)[[2]][[1]]*nrow(BigMac2003)
```

```
## [1] 1.7709
```

For Moscow, because we checked it without looking at the data– we can use the unadjusted P value.

```
test_all <- outlierTest(hyp_fit, cutoff = 1*nrow(BigMac2003), n.max = nrow(BigMac2003), order = TRUE)
test_all$p[5]
```

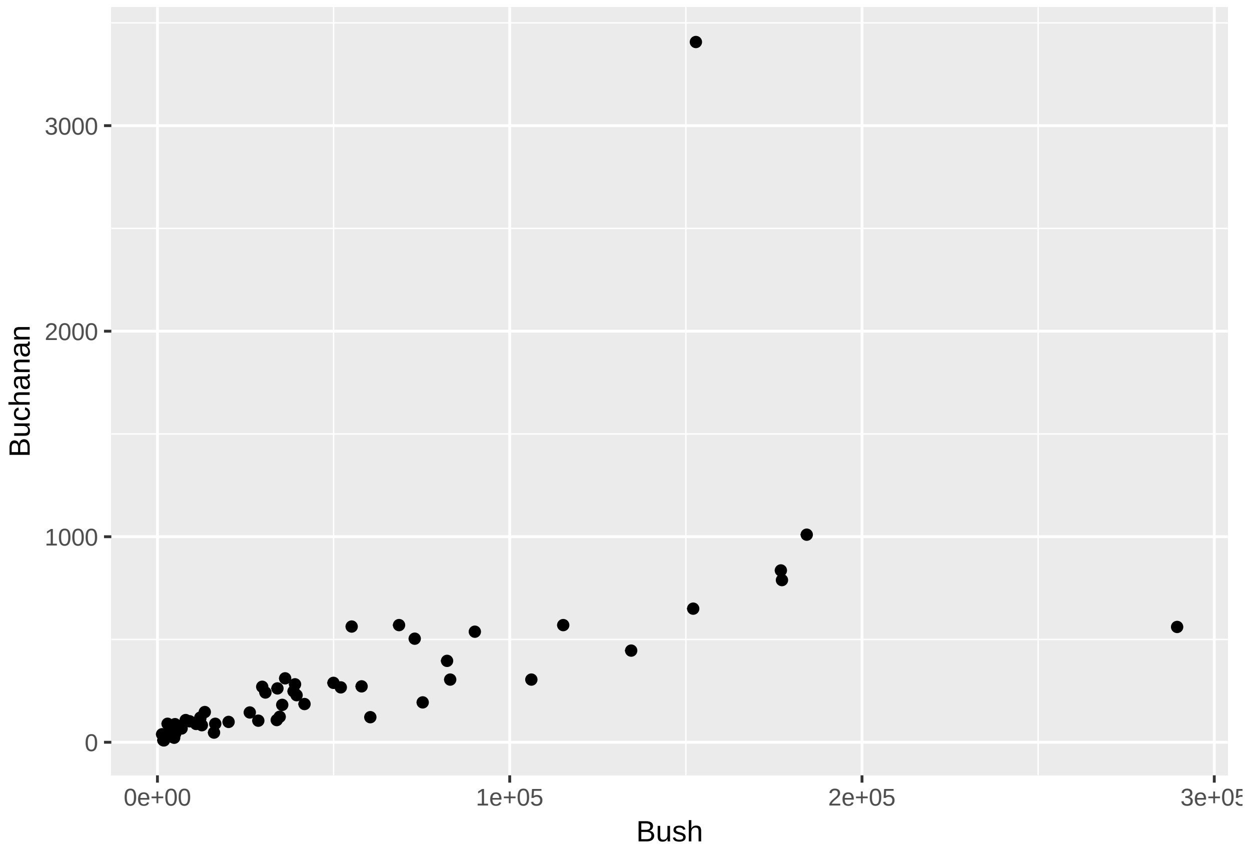
```
##     Moscow
## 0.04965696
```

9.6

Palm Beach is is an outlier justified by the Bonferonni P. Dade is also suspicious but we can't state that it is an outlier via this method. Had we chosen Dade before looking at the data for a specific reason, it might be considered an outlier when it is not because we generated it from the data, which forces us to consider its adjusted p value.

The maximum Cook's distance is over 2. This is a very high Cook's distance, which measures the influence specific points have in the model. It can be thought of as the effect of deleting a point, and is caused by either by being a strong outlier, exhibiting high leverage, or a combination of both.

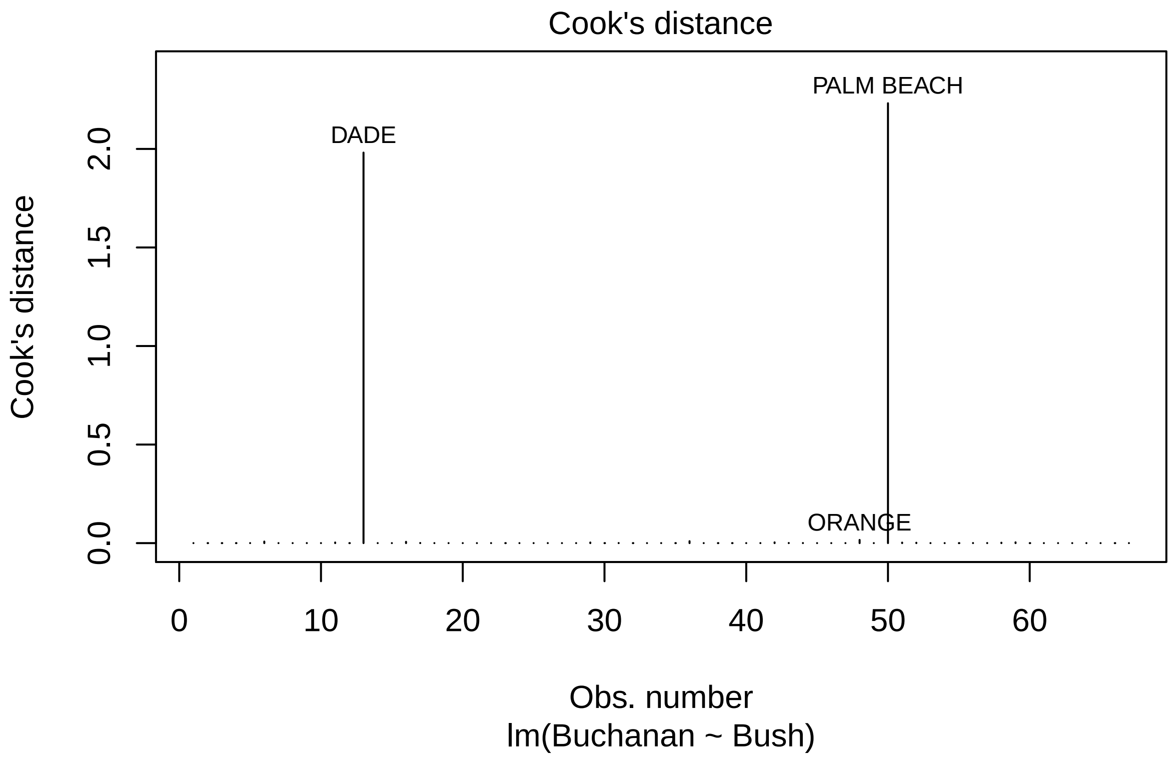
```
ggplot(florida,aes(y=Buchanan,x=Bush)) + geom_point()
```



```
fl_fit <- lm(Buchanan~Bush,data=florida)
outlierTest(fl_fit,cutoff = 2)
```

```
##              rstudent unadjusted p-value Bonferonni p
## PALM BEACH 24.080144         8.6246e-34   5.7785e-32
## DADE       -3.280922         1.6772e-03   1.1237e-01
```

```
plot(fl_fit,which=4)
```



After Transformation, we still find Palm Beach to be an outlier. The maximum cook's distance is about .3.

```
fl_fit_transformed <- lm(log(Buchanan) ~ log(Bush),data=florida)
outlierTest(fl_fit_transformed)
```

```
##            rstudent unadjusted p-value Bonferonni p
## PALM BEACH 4.066282         0.00013325    0.0089278
```

```
outlierTest(fl_fit_transformed)
```

```
##            rstudent unadjusted p-value Bonferonni p
## PALM BEACH 4.066282         0.00013325    0.0089278
```

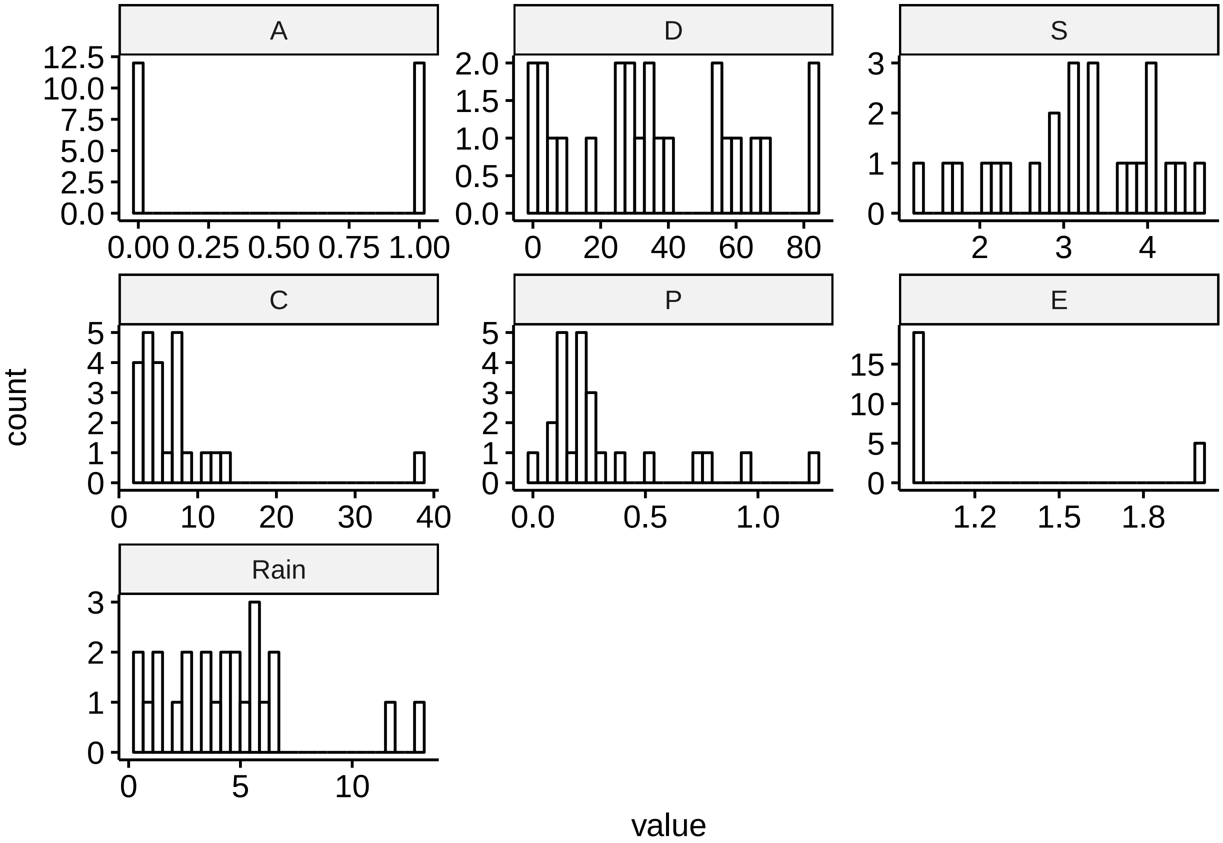
9.7

First, let's look at a histogram of each of our values.

```
cloud_melt <- melt(cloud)
```

```
## Using  as id variables
```

```
gghistogram(cloud_melt,x="value") + facet_wrap(~variable,scales = "free")
```



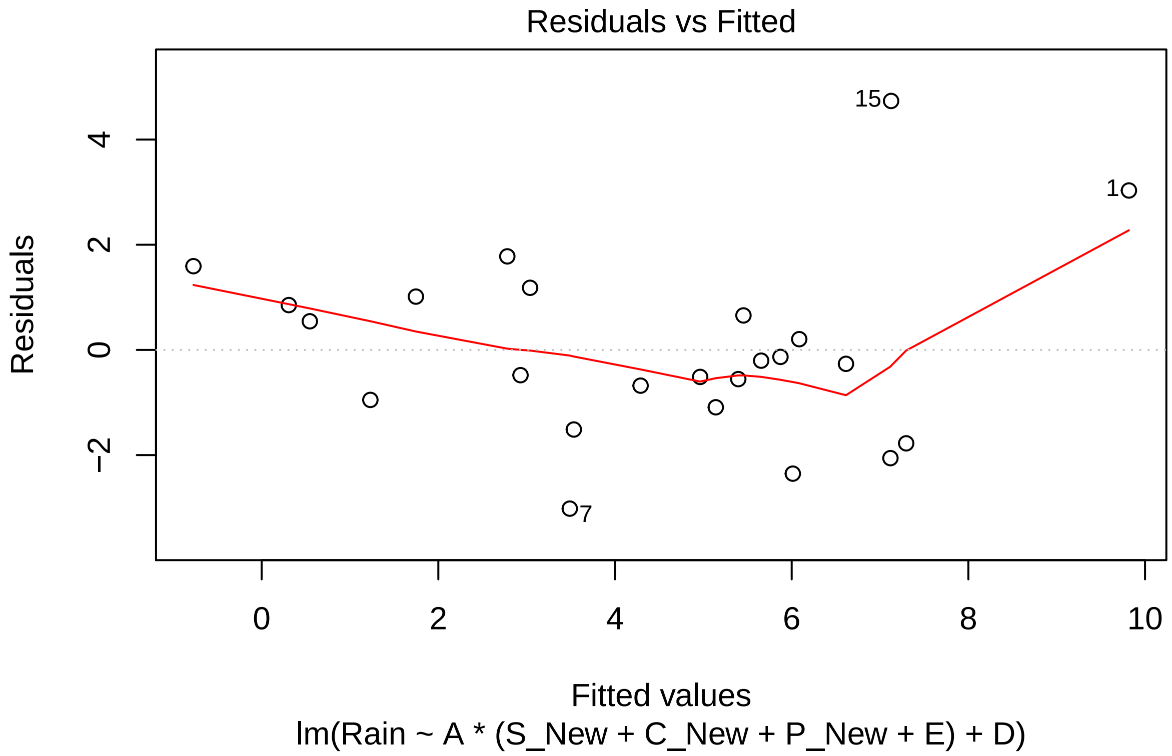
There's considerible non-normal data, so we're going to PowerTransform the non-binary values of our dataset.

```
pt_result <- summary(powerTransform(cbind(cloud$S,cloud$C,cloud$P) ~ 1))$result

cloud$S_New <- pt(cloud$S,pt_result[1])
cloud$C_New <- pt(cloud$C,pt_result[2])
cloud$P_New <- pt(cloud$P,pt_result[3])
```

A logical first principles model is to interact whether or not we seeded with each of the various underlying features that make the record more or less prone to Rainfall, such as its previous rainfall, the type of clouds, and its suitability. We will fit this type of model first, and examine it.

```
fit <- lm(Rain ~ A * (S_New + C_New + P_New + E) + D,data=cloud)
plot(fit,which=1)
```



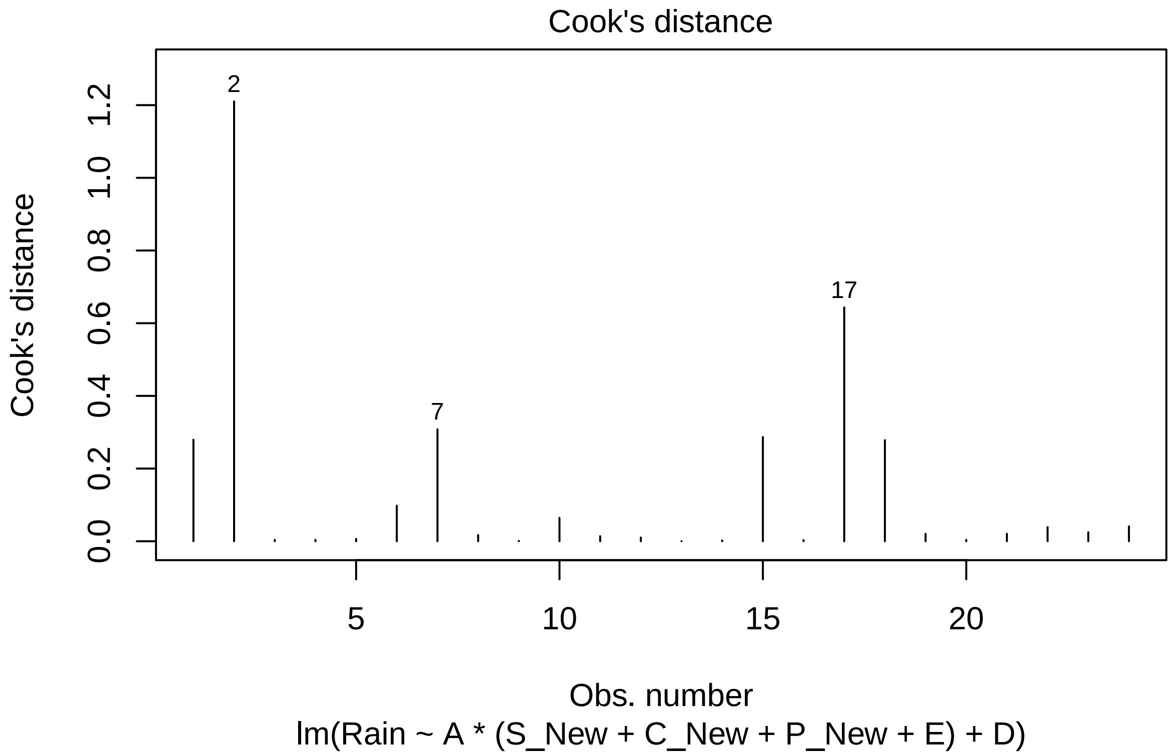
We see what looks like a U shape to our residuals, but its not very strong. But it looks like we are primarily seeing that because of one data point. However, with a Bonferonni P of .13, we can't classify it as an outlier.

```
outlierTest(fit)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 15 3.372105          0.0055486      0.13317
```

We can see that this point has alot of influence. It looks to have a very large amount of cloud cover and Prewetness (previous rainfall). Its three times the amount of cloud cover we saw anywhere else in the dataset. Our next step would be to go look at the data and see if there is a logical reason this happened and whether or not it represents an outlier apart from its results in the data.

```
plot(fit,which=4)
```


Cook's distance
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2
7
17
5
10
15
20
Obs. number
lm(Rain ~ A * (S_New + C_New + P_New + E) + D)