NTHU STAT 5410 - Linear Models

Assignment 6 Report

105061110 周柏宇

1.

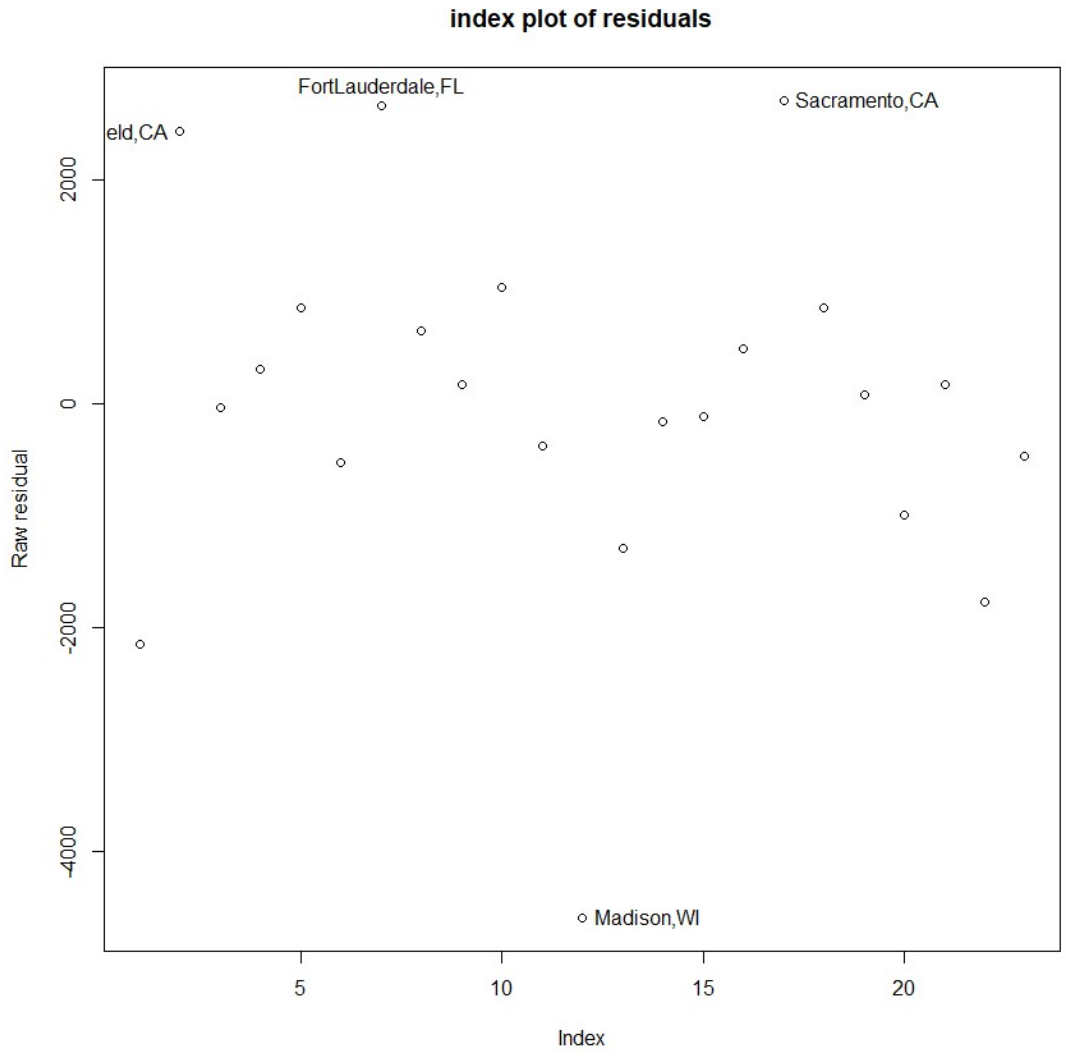
```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear
_models/hw6/crime.txt", header=T)
> name <- gala[,1]
> violent <- gala[,2]
> property <- gala[,3]
> population <- gala[,4]
```

Regress property crime rate on population

```
> fit1 <- lm(property ~ population)
```

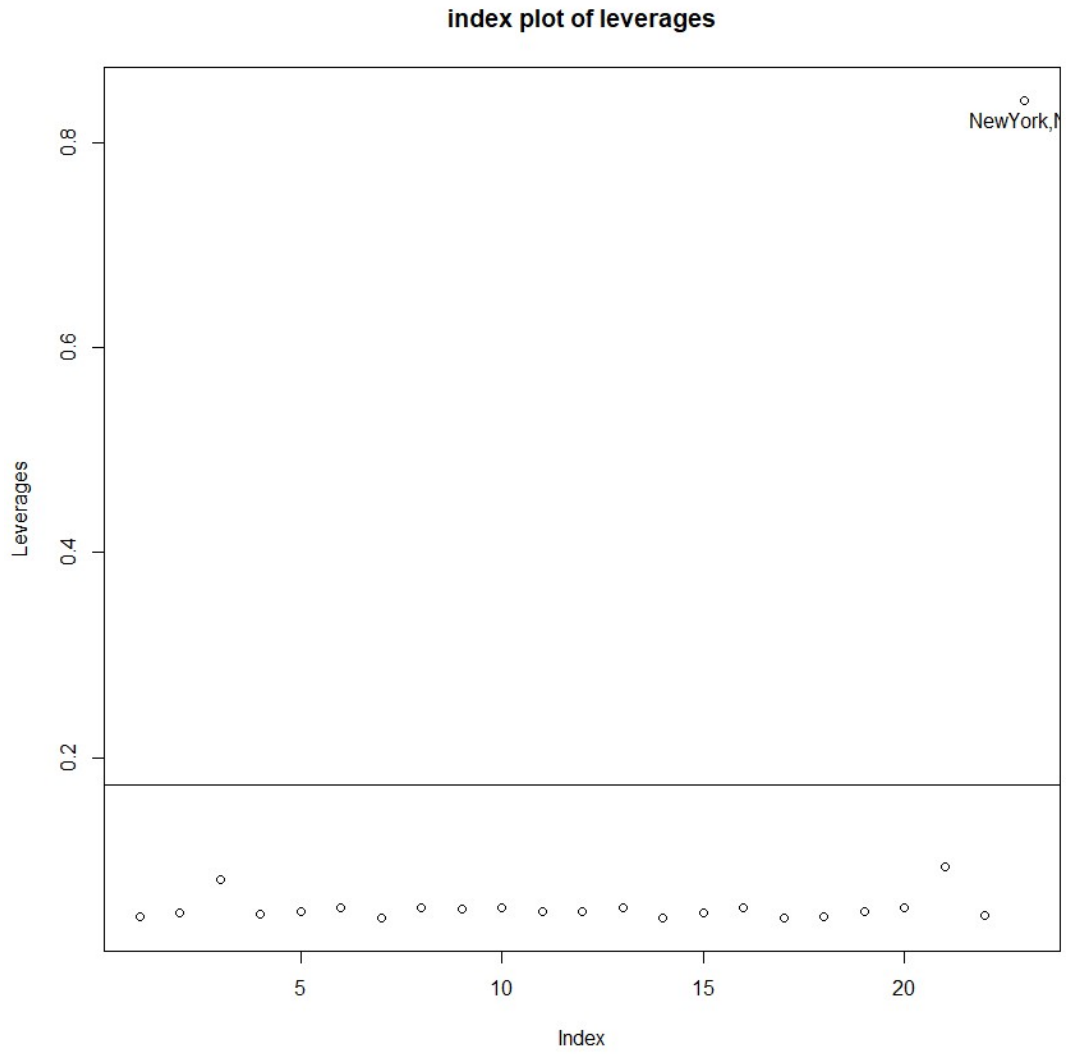
Raw residual plot against index

```
> plot(fit1$res, ylab="Raw residual", main="index plot
 of residuals")
> identify(1:23,fit1$res, name)
```



Leverage plot against index

```
> x1 <- model.matrix(fit1)
> lev1 <- hat(x1)
> plot(lev1, ylab="Leverages",main="index plot of leve
rages")
> abline(h=2*2/23)
> identify(1:23,lev1, name)
```

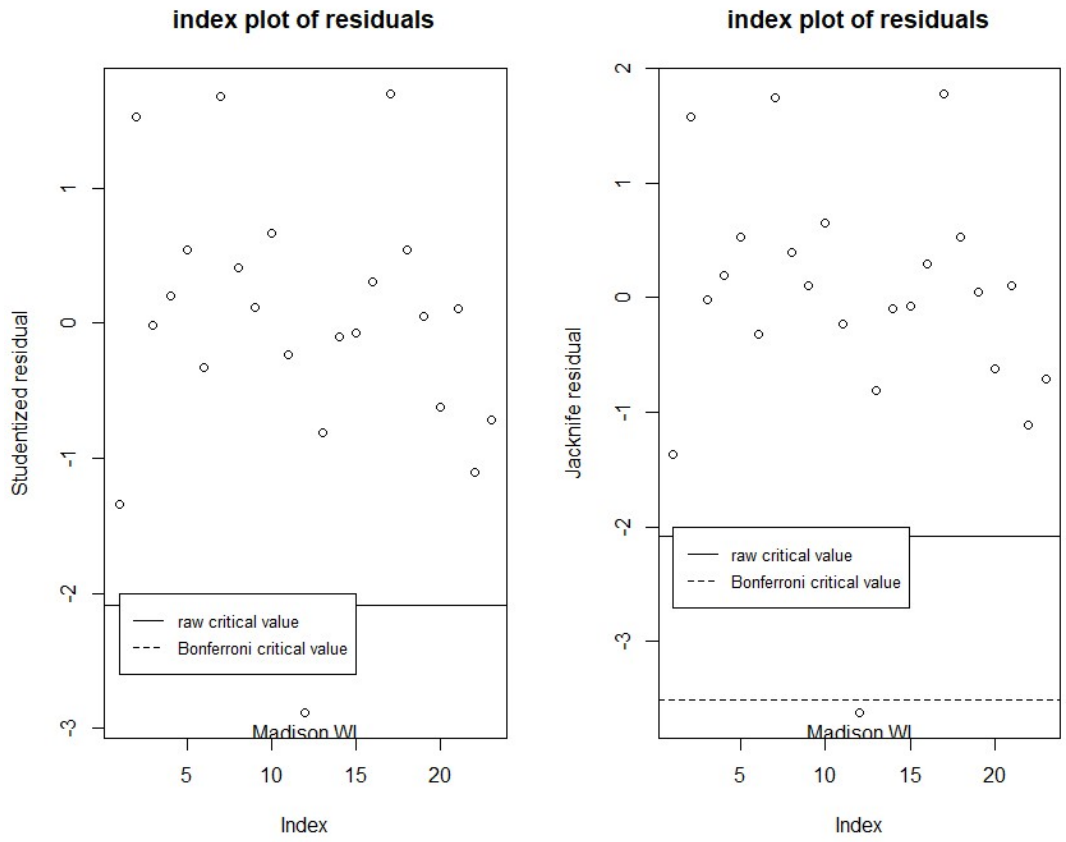



```
> names(lev1) <- name
> lev1[lev1 > 2*2/23]
NewYork,NY
 0.8416482
```

Studentized residual and Jacknife residual plot against index

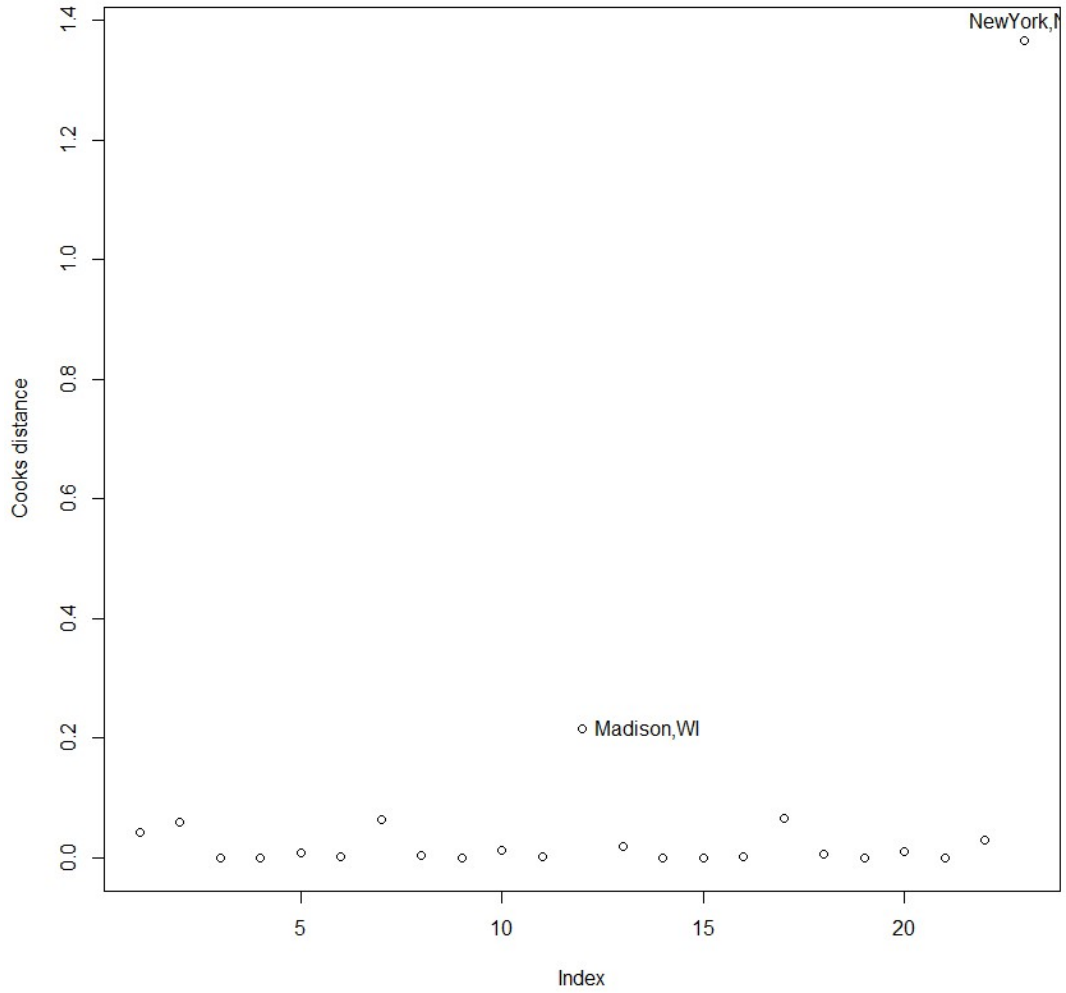
```
> par(mfrow=c(1,2))
> plot(stud1, ylab="Studentized residual", main="index
 plot of residuals");
> abline(cv1, 0, lty=1);abline(-cv1, 0, lty=1);
> abline(cvBF1, 0, lty=2);abline(-cvBF1, 0, lty=2);
> legend(1, -2, legend=c("raw critical value", "Bonferr
oni critical value"), lty=1:2, cex=0.8)
> identify(1:23,stud1, name)
```

```
> plot(jack1, ylab="Jacknife residual", main="index pl
ot of residuals");
> abline(cv1, 0, lty=1);abline(-cv1, 0, lty=1);
> abline(cvBF1, 0, lty=2);abline(-cvBF1, 0, lty=2);
> legend(1, -2, legend=c("raw critical value", "Bonferr
oni critical value"), lty=1:2, cex=0.8)
> identify(1:23,jack1, name)
```



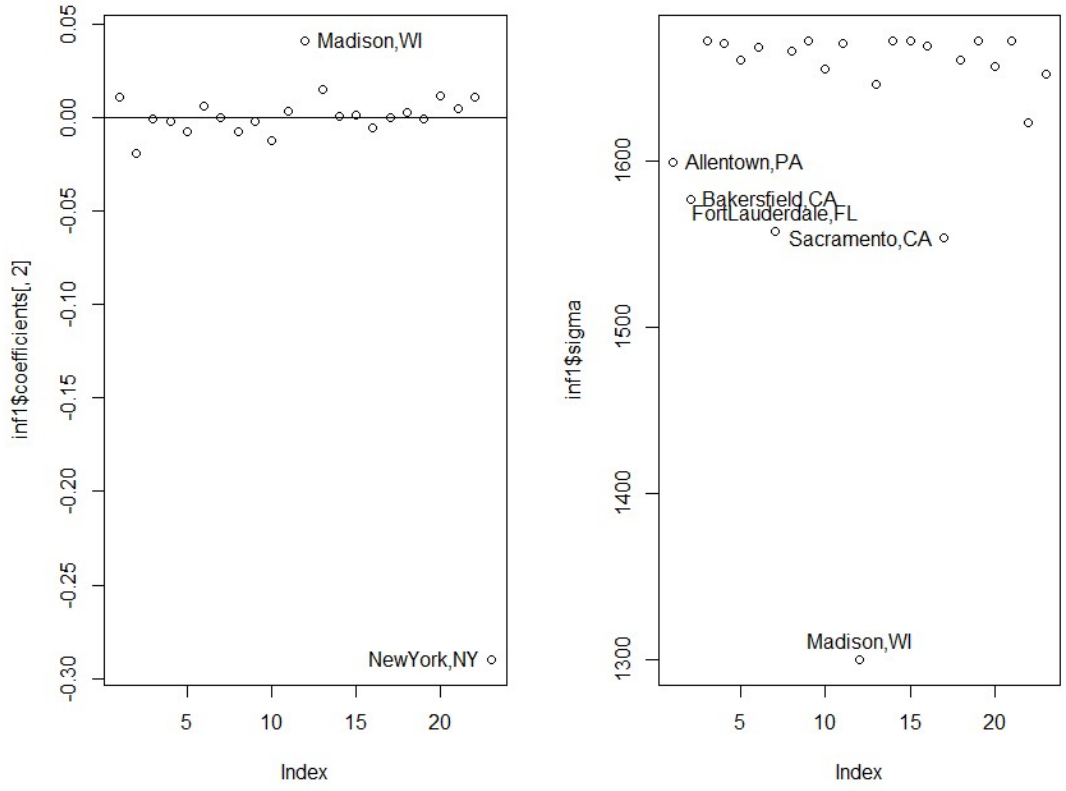
Cook statistics against index

```
> par(mfrow=c(1,1))
> cook1 <- cooks.distance(fit1)
> plot(cook1, ylab="Cooks distance")
> identify(1:23,cook1,name)
```



Influence of leave-out-one coefficients

```
> inf1 <- lm.influence(fit1)
> par(mfrow=c(1,2))
> plot(inf1$coefficients[,2]) # population
> abline(0, 0)
> identify(1:23, inf1$coefficients[,2], name)
> plot(inf1$sigma)
> identify(1:23, inf1$sigma, name)
```



Summary of diagnostics on fit1

	NewYork,NY	Madison,WI	Sacramento, CA	FortLauderdale, FL	Bakersfield, CA	Allentown,PA
large leverage	*					
studentized residual						
jacknife residual		*				
Cook stat.	**	*				
diff. in LO1 coef. (population)	**	*				
change in LO1 sigma		**	*	*	*	*

We would like to fit a model without observations of NewYork,NY and Madison,WI judging by the diagnostics, the new model will then be:

```
> fit11 <- lm(property ~ population, subset=(name!="New
York,NY" & name!="Madison,WI"))
> summary(fit11)

Call:
lm(formula = property ~ population, subset = (name != "NewYork,NY"
 &
    name != "Madison,WI"))

Residuals:
     Min       1Q   Median       3Q      Max
-2380.96  -632.54   -91.44   509.00  2413.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5323.0265   375.4223  14.179 1.48e-11 ***
population     0.3462     0.3591   0.964    0.347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1319 on 19 degrees of freedom
Multiple R-squared:  0.04663,      Adjusted R-squared:  -0.003546
F-statistic: 0.9293 on 1 and 19 DF,  p-value: 0.3471
```

Compared with the original

```
> summary(fit1)

Call:
lm(formula = property ~ population)

Residuals:
    Min      1Q  Median      3Q     Max
-4595.2  -493.9    81.8   753.6  2713.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5198.4645   387.2227  13.425 8.95e-12 ***
population     0.1728     0.1802   0.959    0.348
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1633 on 21 degrees of freedom
Multiple R-squared:  0.04196,      Adjusted R-squared:  -0.003662
F-statistic: 0.9197 on 1 and 21 DF,  p-value: 0.3485
```

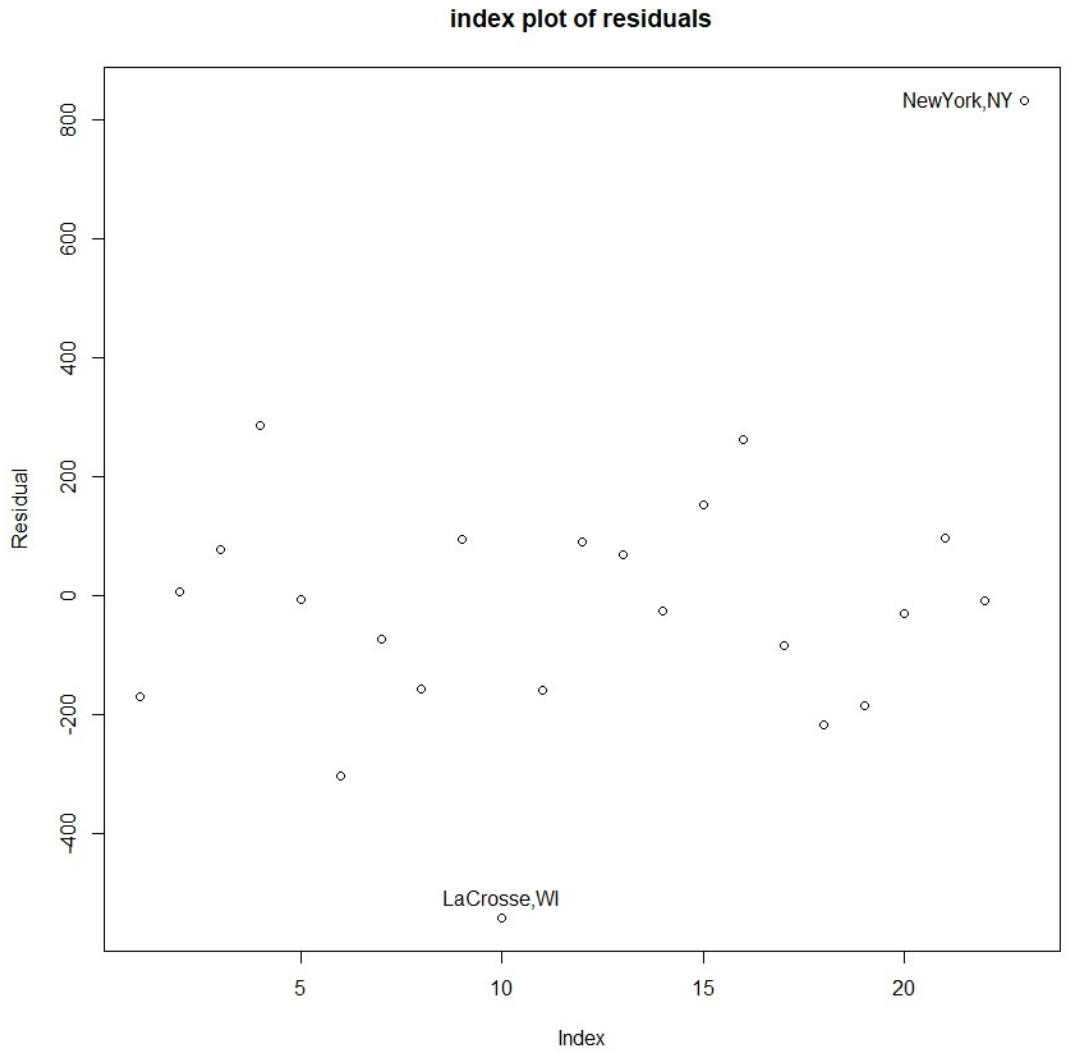
The coefficient of the predictor “population” has about 100.35% of difference and the range of residual reduces.

Regress violent crime rate on property crime rate

```
> fit2 <- lm(violent ~ property)
```

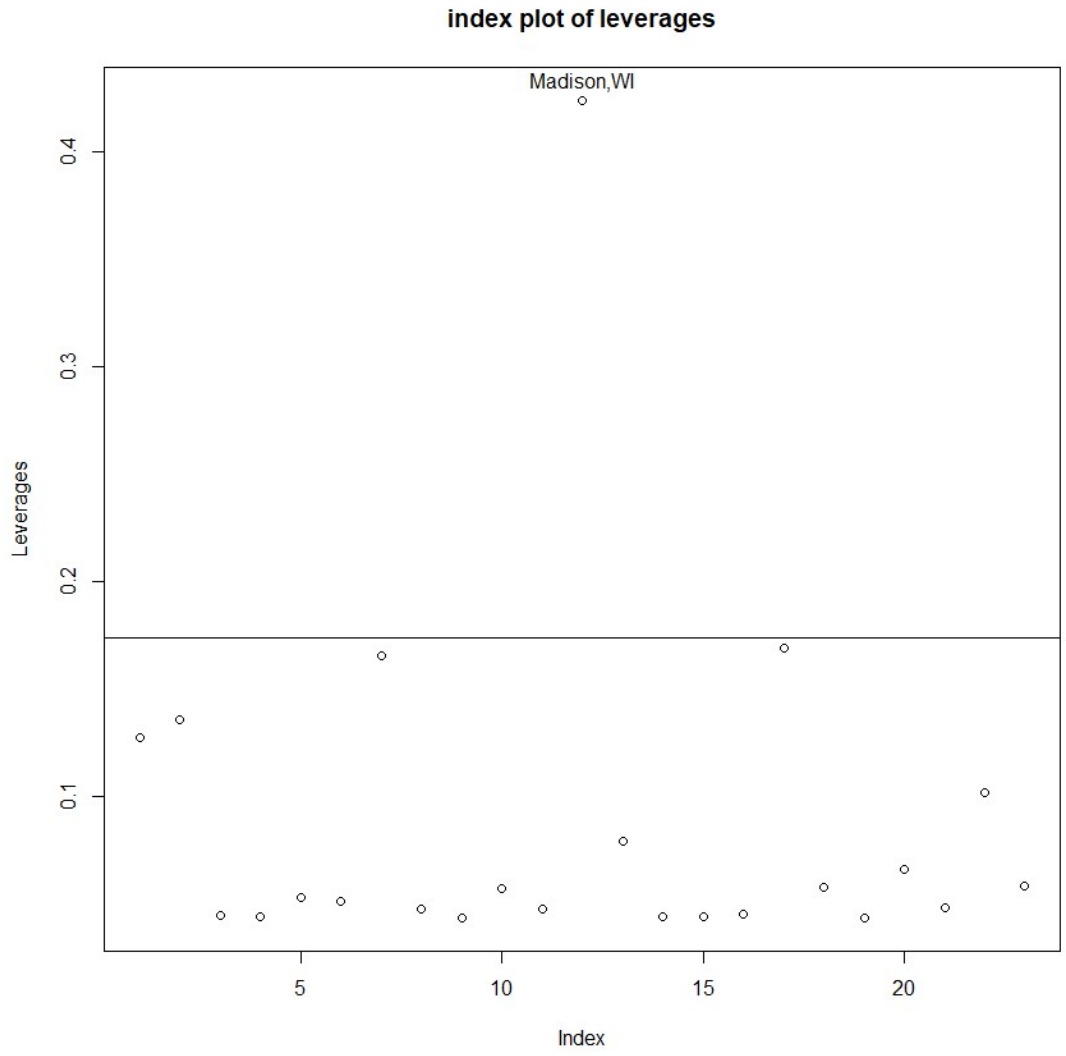
Raw residual plot against index

```
> plot(fit2$res, ylab="Residual", main="index plot of r
esiduals")
> identify(1:23,fit2$res, name)
```



Leverage plot against index

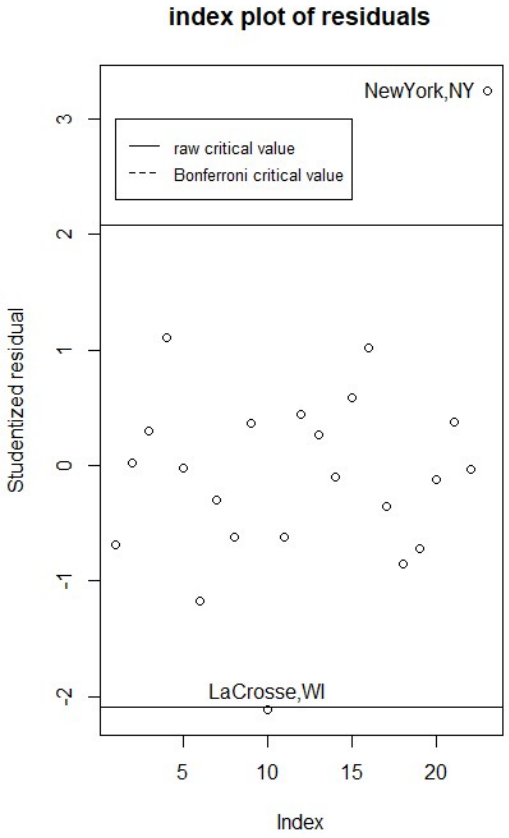
```
> x2 <- model.matrix(fit2)
> lev2 <- hat(x2)
> plot(lev2, ylab="Leverages",main="index plot of leve
rages")
> abline(h=2*2/23)
> names(lev2) <- name
> lev2[lev2 > 2*2/23]
Madison,WI
 0.4240726
```

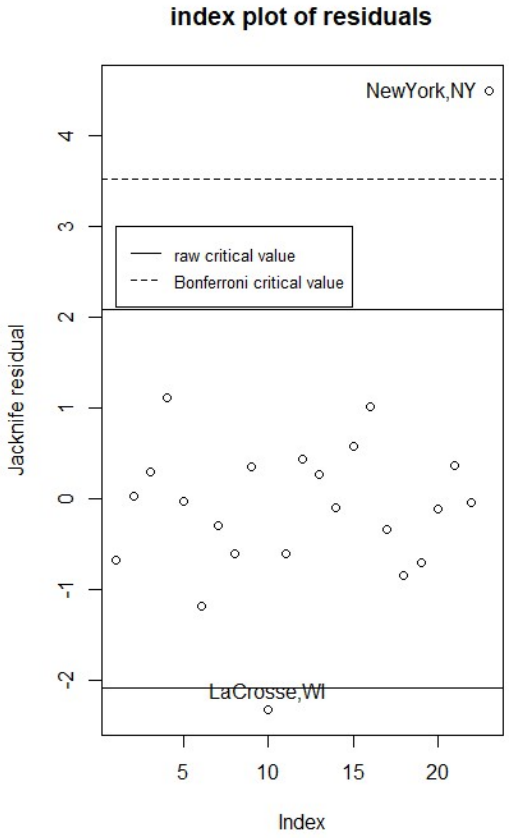


Studentized residual and Jacknife residual plot against index

```
> par(mfrow=c(1,2))
> plot(stud2, ylab="Studentized residual", main="index
 plot of residuals");
> abline(cv2, 0, lty=1);abline(-cv2, 0, lty=1);
> abline(cvBF2, 0, lty=2);abline(-cvBF2, 0, lty=2);
> legend(1, 3, legend=c("raw critical value", "Bonferro
ni critical value"), lty=1:2, cex=0.8)
> identify(1:23,stud2, name)
> plot(jack2, ylab="Jacknife residual", main="index pl
ot of residuals");
> abline(cv2, 0, lty=1);abline(-cv2, 0, lty=1);
> abline(cvBF2, 0, lty=2);abline(-cvBF2, 0, lty=2);
```

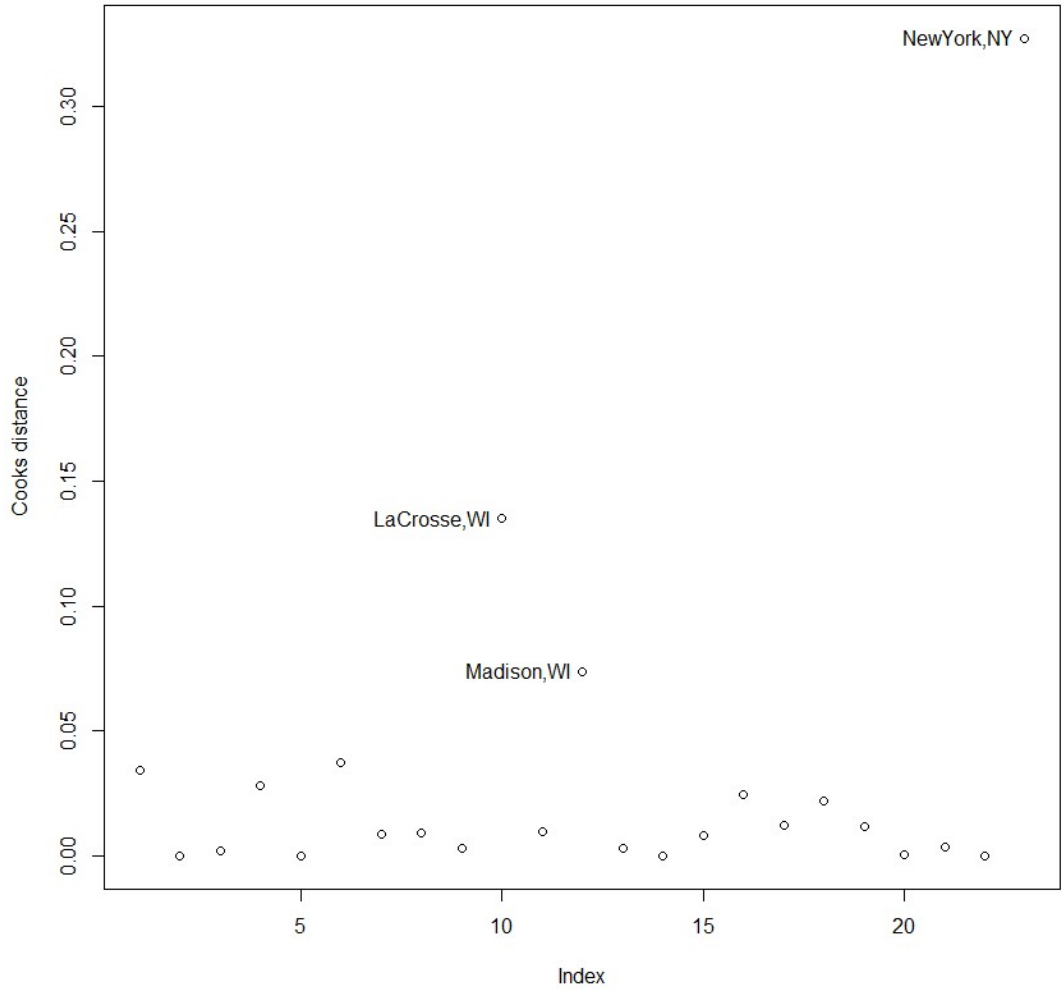
```
> legend(1, 3, legend=c("raw critical value", "Bonferro
ni critical value"), lty=1:2, cex=0.8)
> identify(1:23,jack2, name)
```





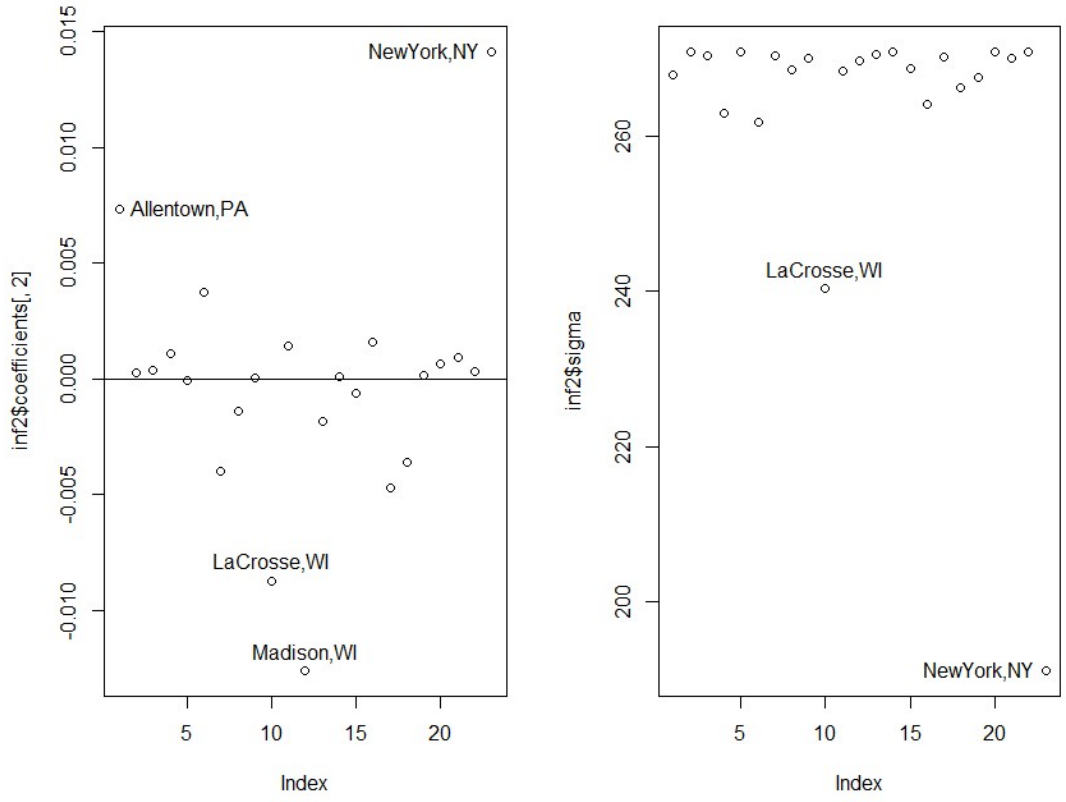
Cook statistics against index

```
> cook2 <- cooks.distance(fit2)
> plot(cook2, ylab="Cooks distance")
> identify(1:23,cook2,name)
```



Influence of leave-out-one coefficients

```
> par(mfrow=c(1,2))
> plot(inf2$coefficients[,2]) # property
> abline(0, 0)
> identify(1:23, inf2$coefficients[,2], name)
> plot(inf2$sigma)
> identify(1:23, inf2$sigma, name)
> par(mfrow=c(1,1))
```



Summary of diagnostics on fit2

	Madison,WI	NewYork,NY	LaCrosse,WI	Allentown,PA
large leverage	**			
studentized residual		*	*	
jacknife residual		**	*	
Cook stat.	*	**	*	
diff. in LO1 coef. (property)	*	**	*	*
change in LO1 sigma	*	**		

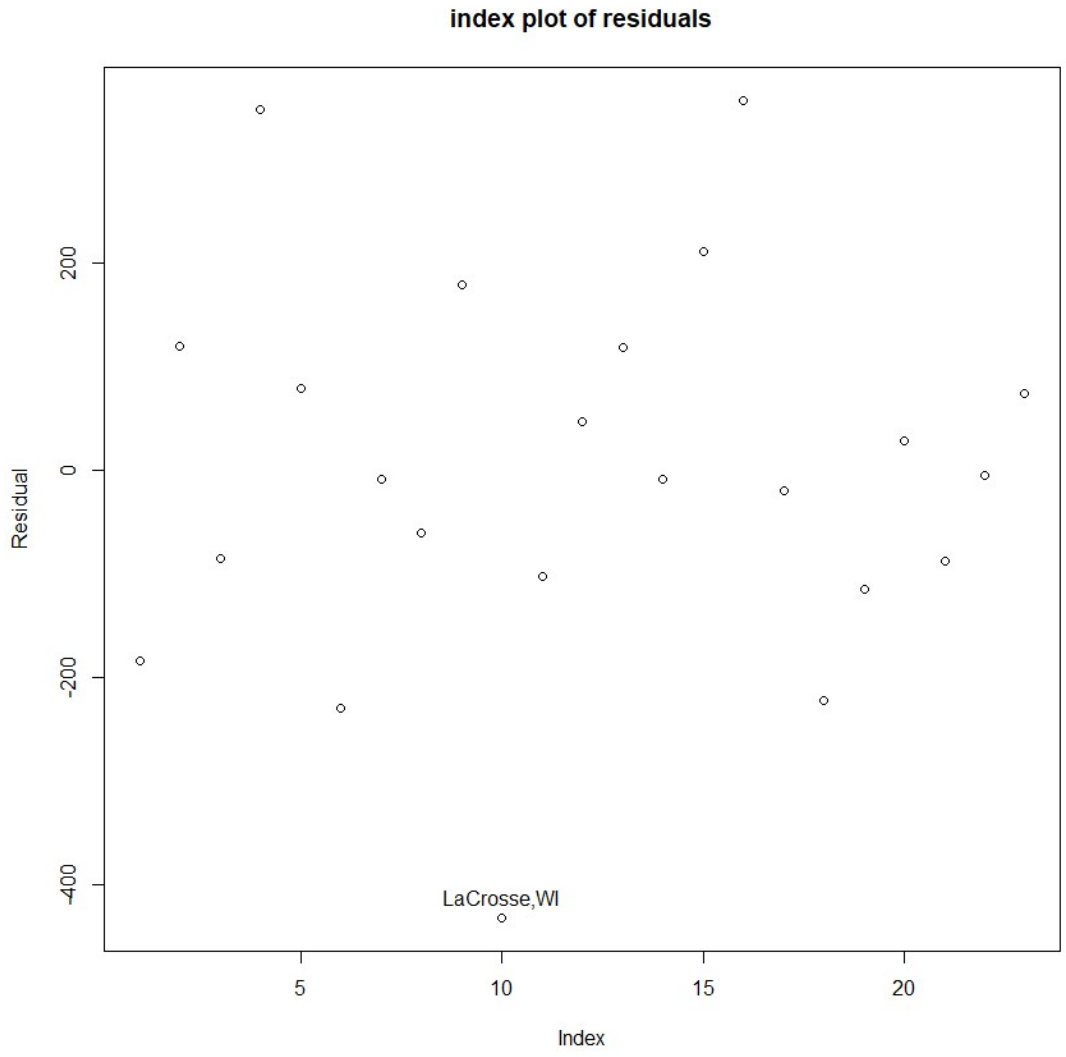
With the result of diagnostics for both fit1 and fit2, we have strong confidence to state that the observations of “NewYork,NY” and “Madison,WI” are outliers to this dataset.

Regression of violent crime rate against property crime rate and population

```
> fit3 <- lm(violent ~ property + population)
```

Raw residual plot against index

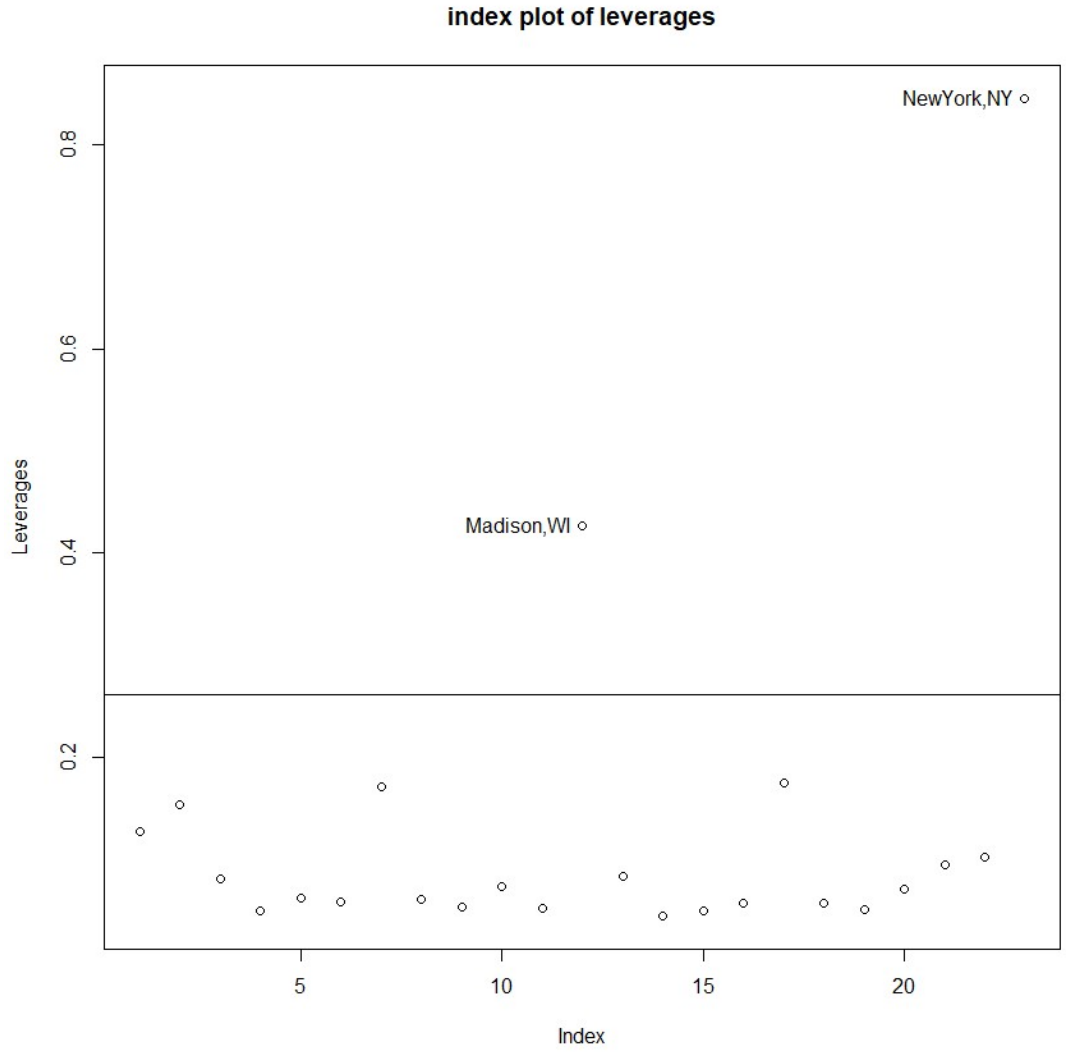
```
> plot(fit3$res, ylab="Residual", main="index plot of r
esiduals")
> identify(1:23,fit3$res, name)
```



Leverage plot against index

```
> plot(lev3, ylab="Leverages",main="index plot of leve
rages")
> abline(h=2*3/23)
> identify(1:23, lev3, name)
> names(lev3) <- name
```

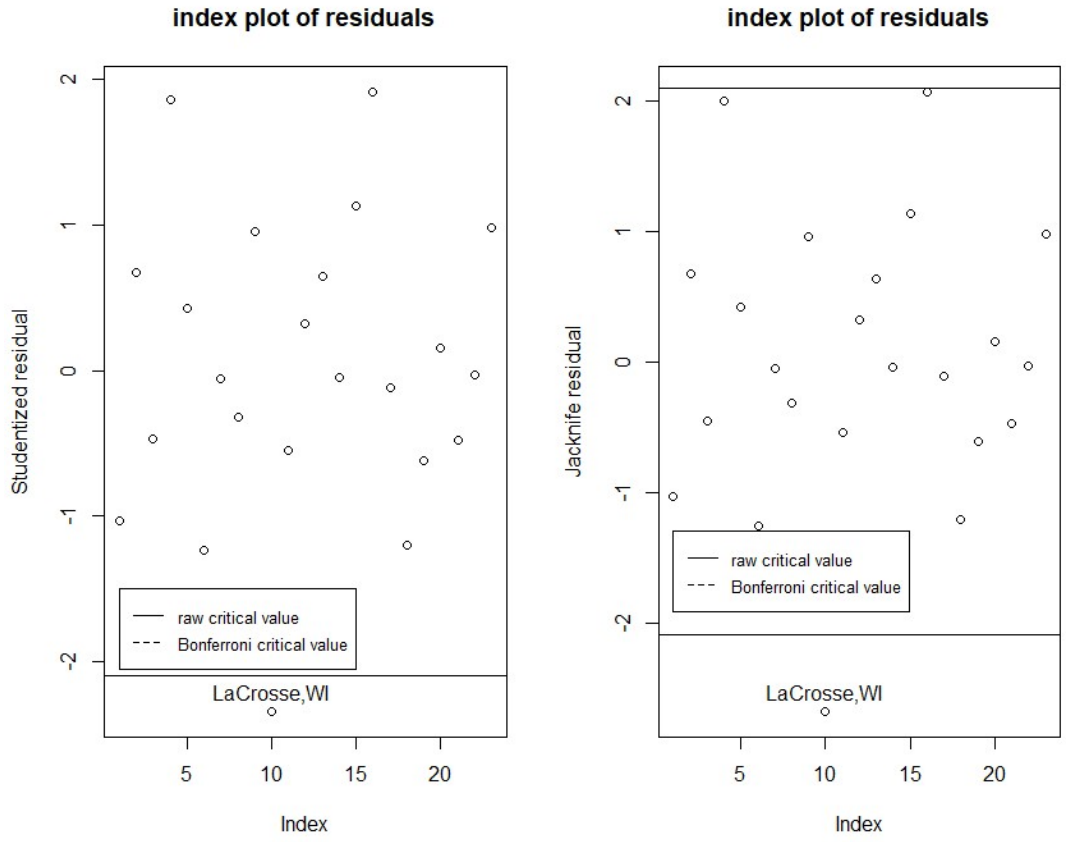
```
> lev3[lev3 > 2*3/23]
Madison,WI NewYork,NY
 0.4265893  0.8455262
```



Studentized residual and Jacknife residual plot against index

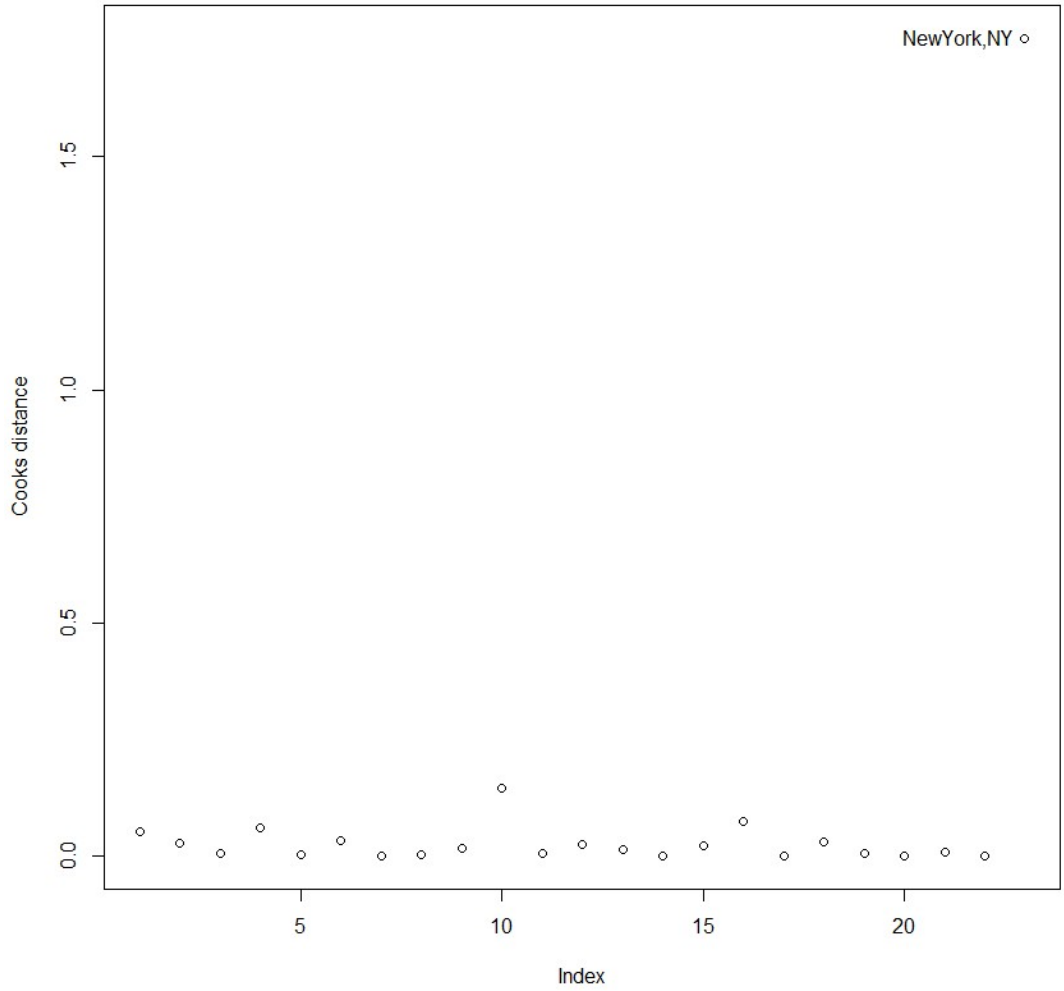
```
> cv3 <- qt(0.05/2, 23-3-1)
> cvBF3 <- qt(0.05/(23*2), 23-3-1)
> stud3 <- rstandard(fit3)
> jack3 <- rstudent(fit3)
> par(mfrow=c(1,2))
> plot(stud3, ylab="Studentized residual", main="index
 plot of residuals");
> abline(cv3, 0, lty=1);abline(-cv3, 0, lty=1);
> abline(cvBF3, 0, lty=2);abline(-cvBF3, 0, lty=2);
```

```
> legend(1, -1.5, legend=c("raw critical value", "Bonfe
rroni critical value"), lty=1:2, cex=0.8)
> identify(1:23,stud3, name)
> plot(jack3, ylab="Jacknife residual", main="index pl
ot of residuals");
> abline(cv3, 0, lty=1);abline(-cv3, 0, lty=1);
> abline(cvBF3, 0, lty=2);abline(-cvBF3, 0, lty=2);
> legend(1, -1.3, legend=c("raw critical value", "Bonfe
rroni critical value"), lty=1:2, cex=0.8)
> identify(1:23,jack3, name)
```



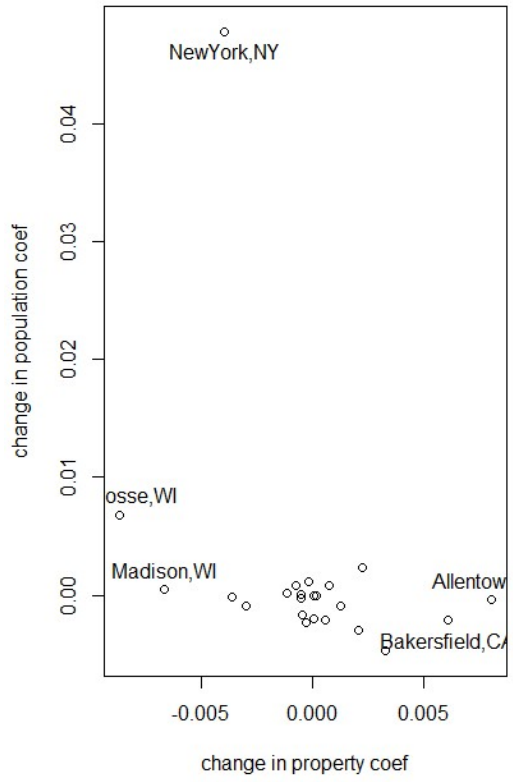
Cook statistics against index

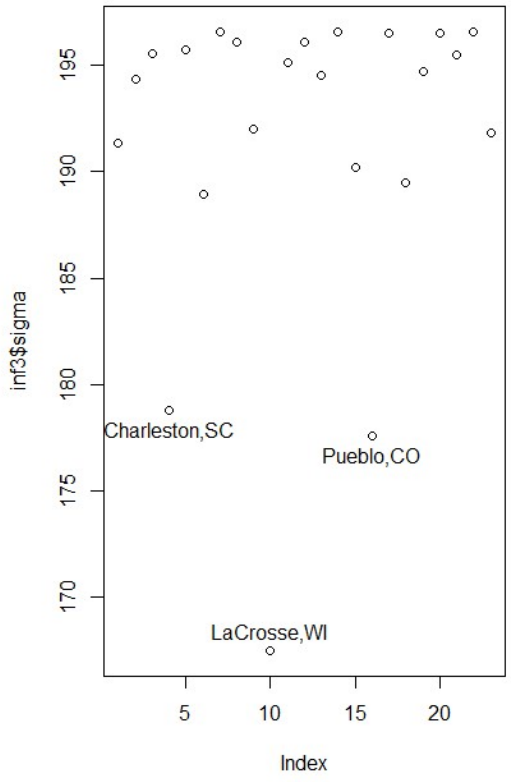
```
> par(mfrow=c(1,1))
> cook3 <- cooks.distance(fit3)
> plot(cook3, ylab="Cooks distance")
> identify(1:23,cook3,name)
```



Influence of leave-out-one coefficients

```
> inf3 <- lm.influence(fit3)
> par(mfrow=c(1,2))
> plot(inf3$coef[,2],inf3$coef[,3],xlab="change in pro
perty coef", ylab="change in population coef")
> identify(inf3$coef[,2],inf3$coef[,3],name)
> plot(inf3$sigma)
> identify(1:23, inf3$sigma, name)
```





Summary of diagnostics on fit3

	Madison,WI	NewYork,NY	LaCrosse,WI	Allentown,PA	Bakersfield,CA
large leverage	*	**			
studentized residual			*		
jacknife residual			*		
Cook stat.		**			
diff. in LO1 coef. (property)	*	*	*	*	*
diff. in LO1 coef. (population)		**	*		
change in LO1 sigma			**		

change in LO1 sigma: Charleston,SC *, Pueblo,CO *

With this model, we may say that the observations of "NewYork,NY" and "LaCrosse,WI" are outliers.

2.

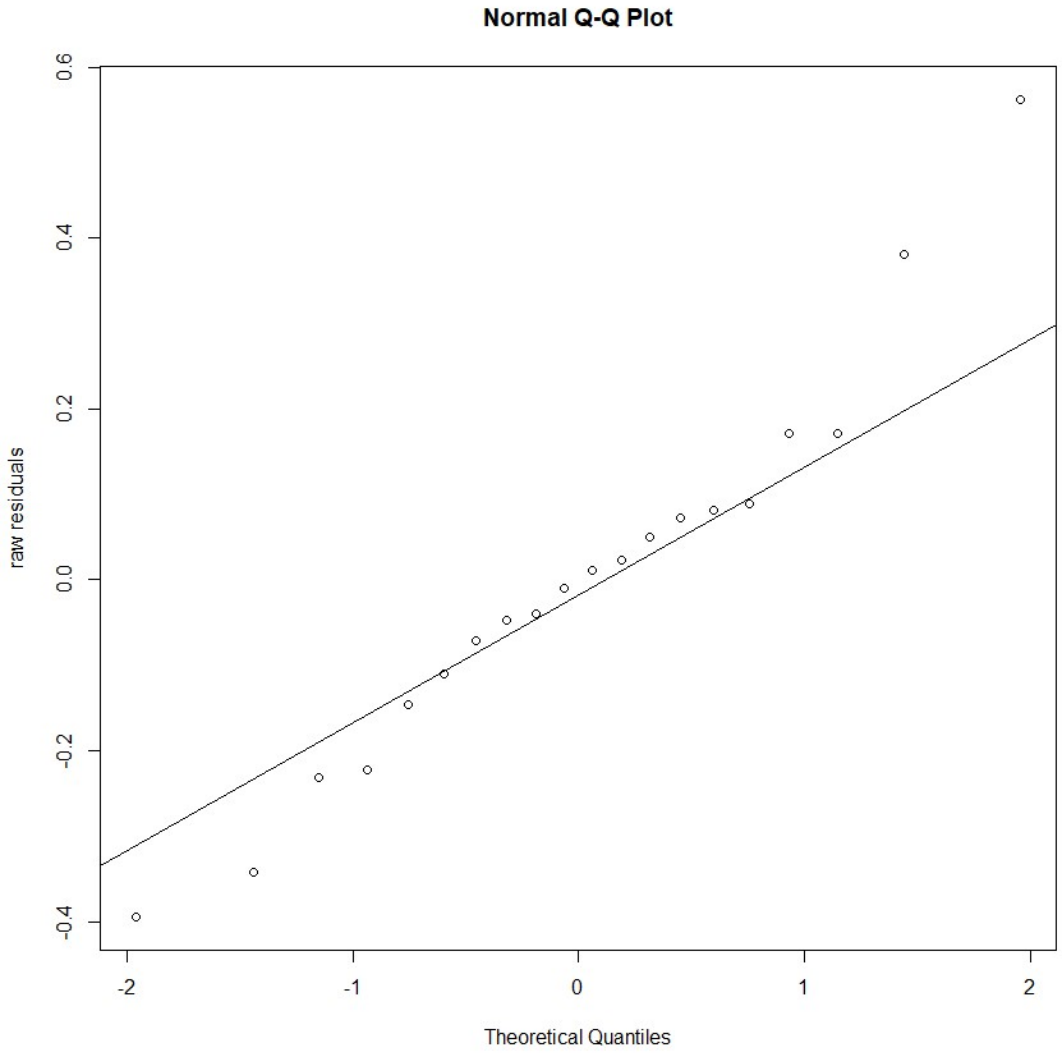
```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear
_models/hw6/E3.7.txt", header=T)
> y <- gala[,7]
> x1 <- gala[,2]
> x2 <- gala[,3]
> x3 <- gala[,4]
> x4 <- gala[,5]
> x5 <- gala[,6]
```

(a)

```
> fit1 <- lm(y ~ x1 + x2 + x3 + x4 + x5)
```

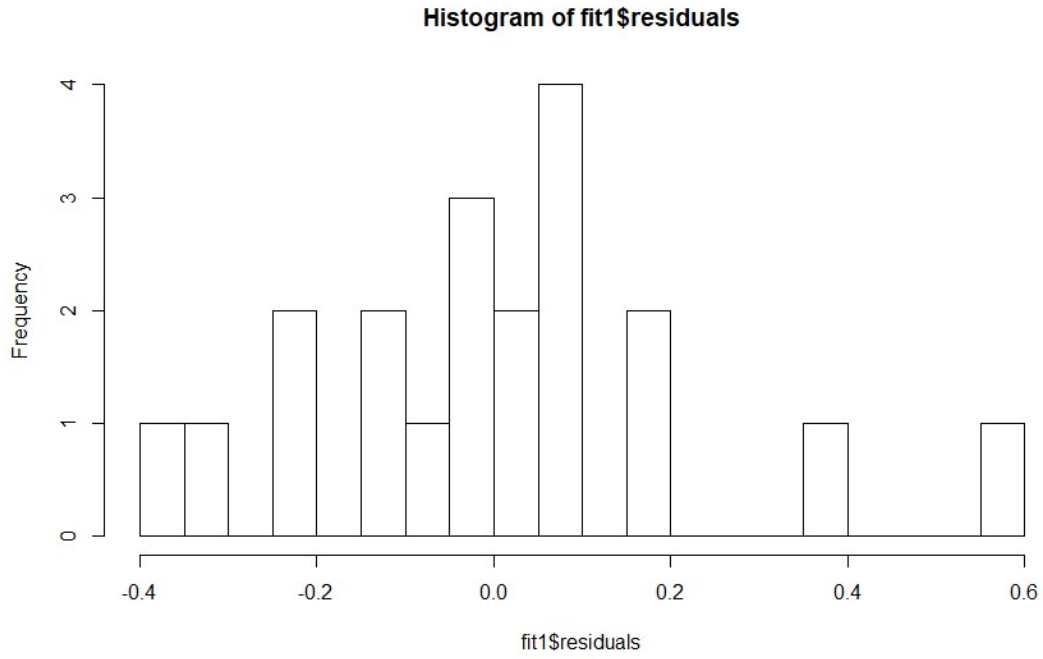
Q-Q plot

```
> qqnorm(fit1$res, ylab="raw residuals")
> qqline(fit1$res)
```



Histogram

```
> hist(fit1$residuals, breaks=20)
```



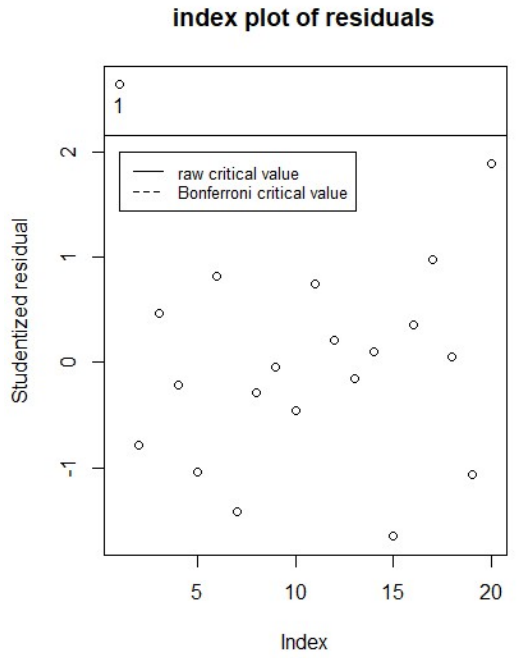
We can see that the low raw residuals are slightly below the line and the high raw residuals are above the line more obviously. This residual distribution is slightly heavy-tailed.

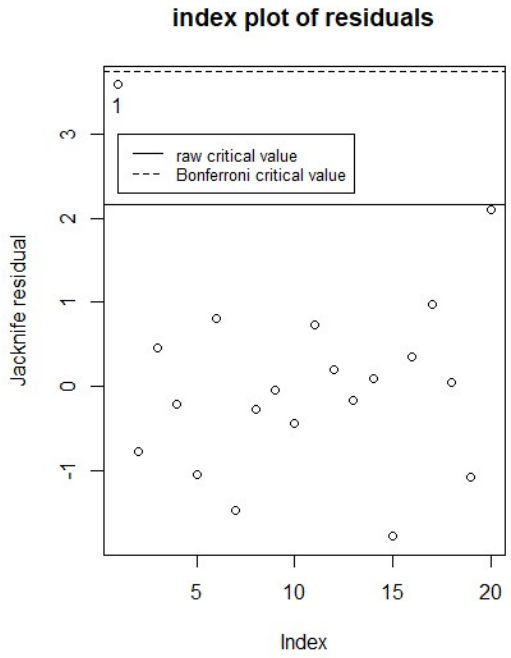
To identify the outlier, we can go through the same procedure in 1.

Laverage

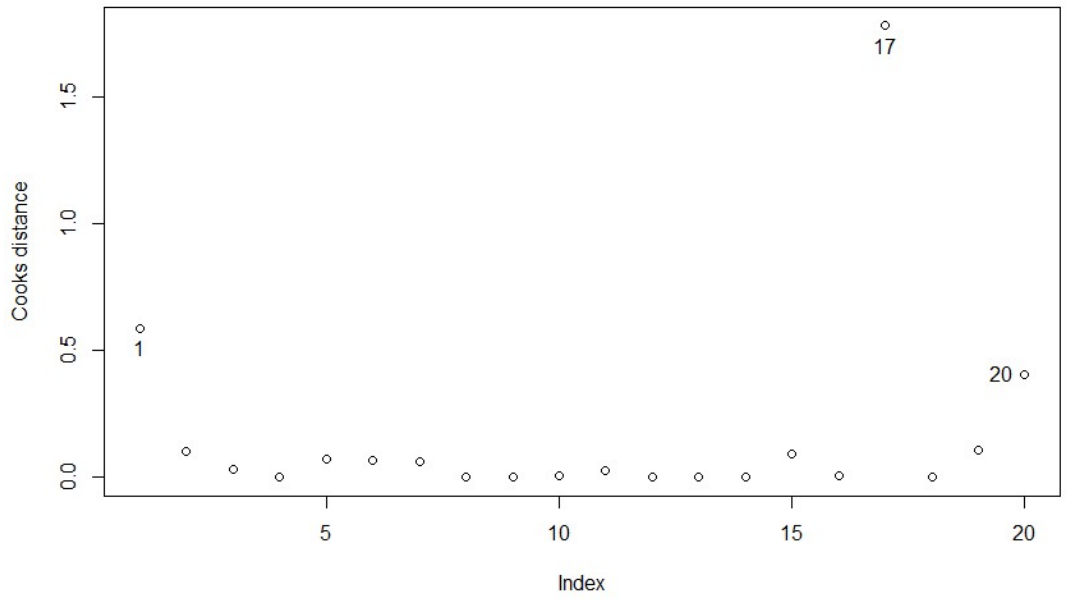
```
> lev1[lev1 > 2*6/20]
   17
0.9182393
```

Studentized residual and Jacknife residual plot against index



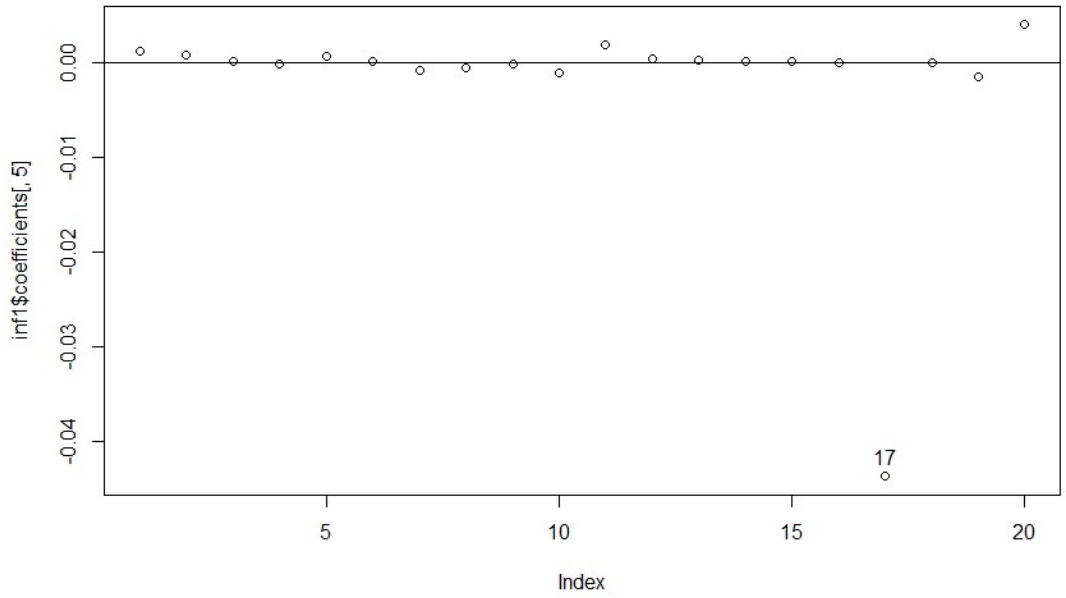


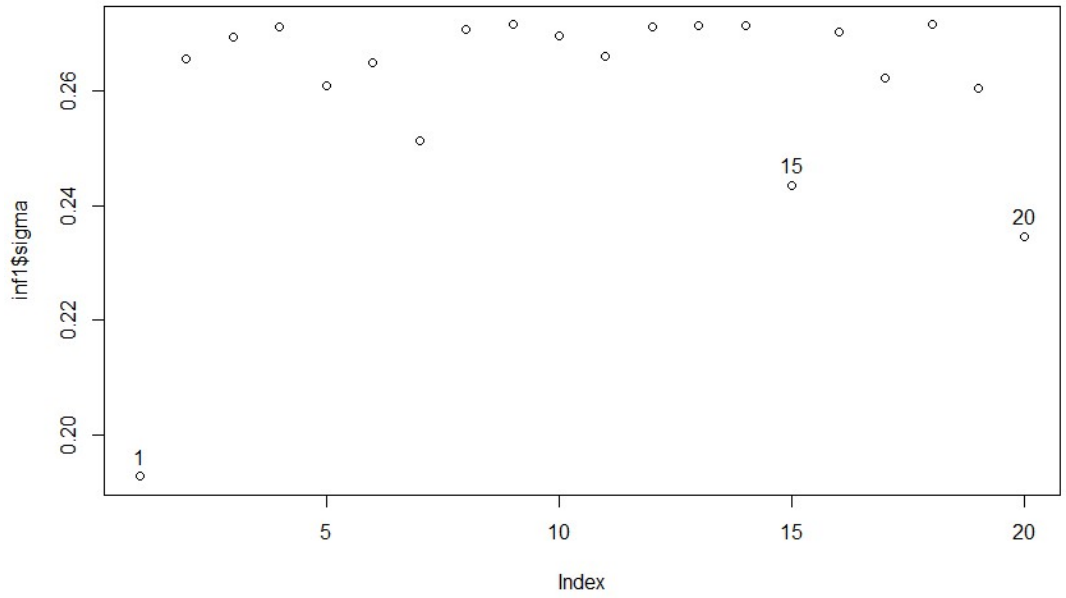
Cook statistics against index



Influence of leave-out-one coefficients

Difference of coefficients for all predictors except “x4” are at a 1e-4 ~ 1e-5 scale, thus omitted.





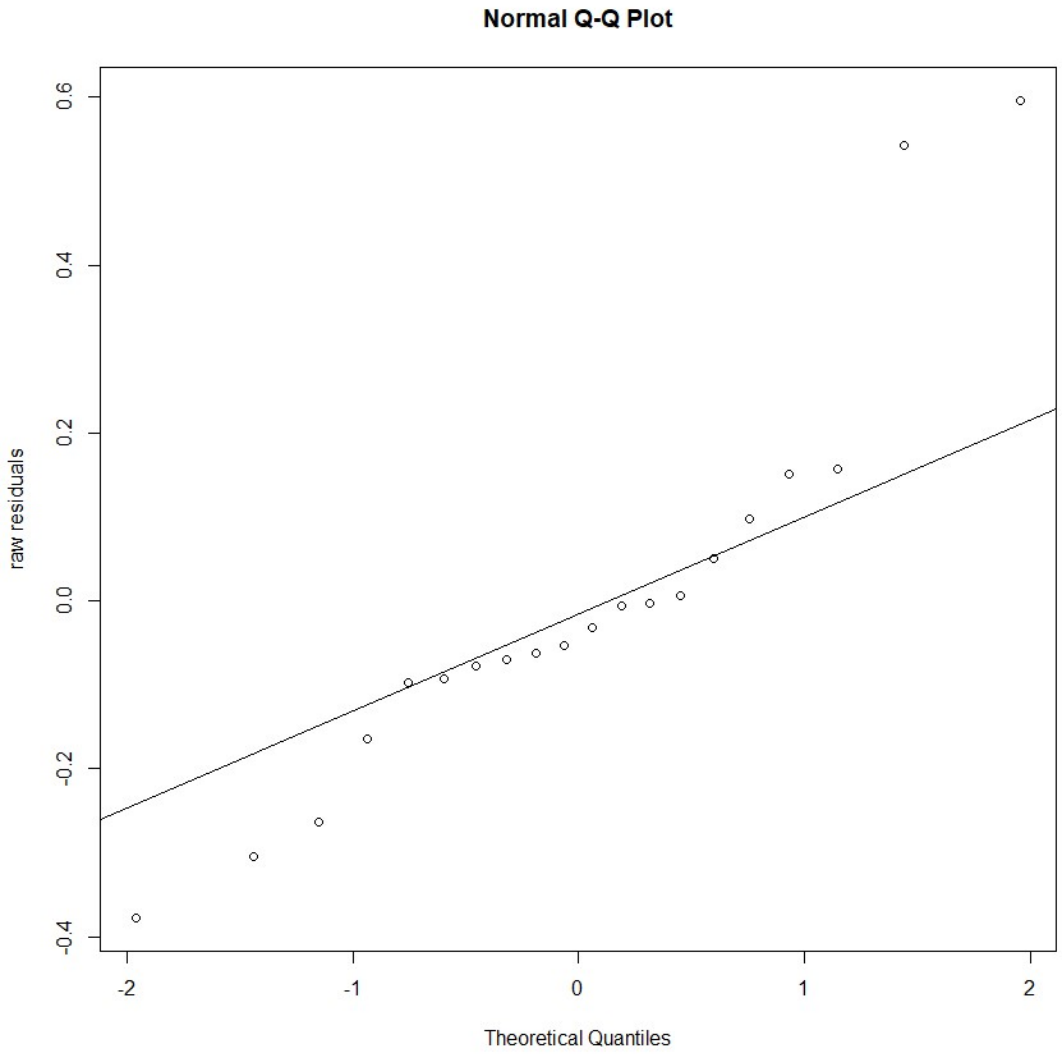
From the result of diagnostics, we may conclude the observation with index 1, 17 and 20 have greater chance of being an outlier.

(b)

```
> fit2 <- lm(y ~ x3 + x5)
```

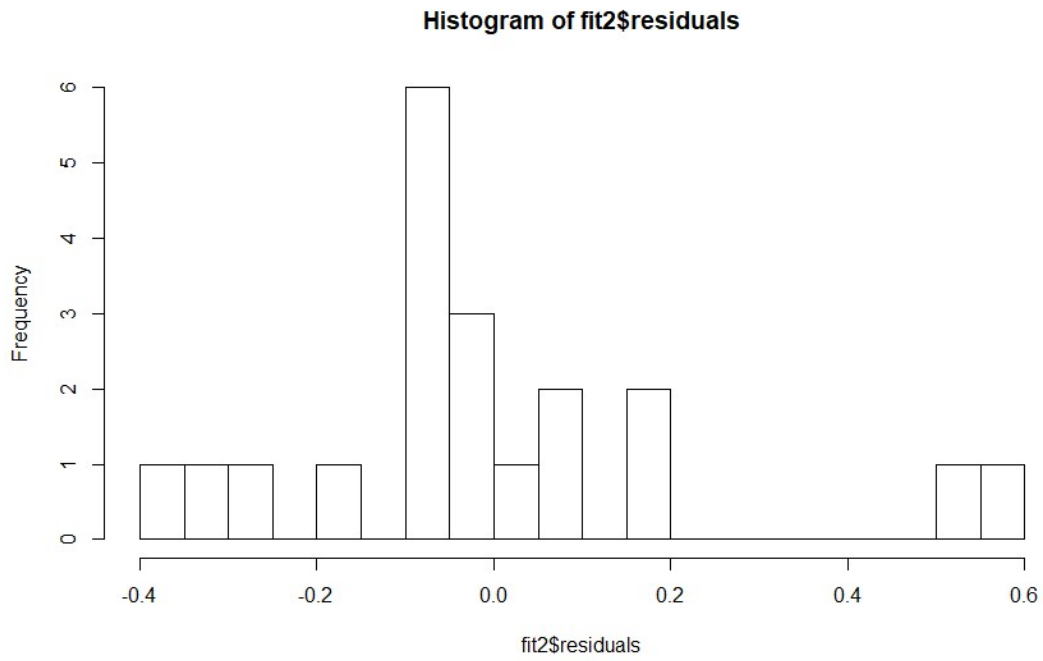
Q-Q plot

```
> qqnorm(fit2$res, ylab="raw residuals")
> qqline(fit2$res)
```



Histogram

```
> hist(fit2$residuals, breaks=20)
```



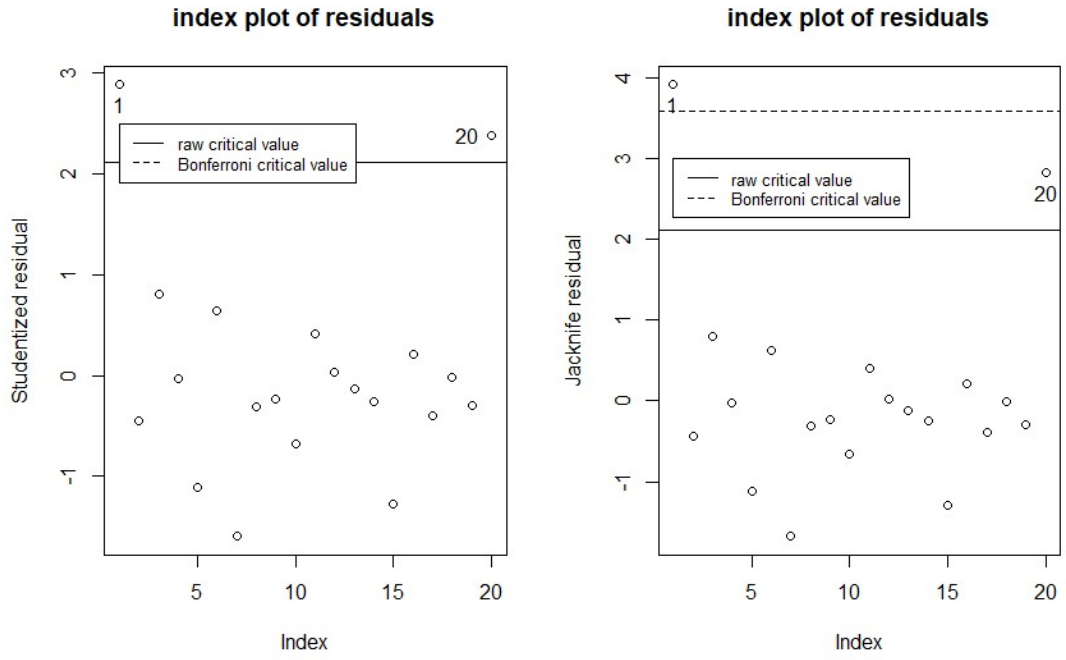
For the distribution of residuals, we can more confidently conclude that it is heavy-tailed from both Q-Q plot and histogram.

Conducting the diagnostics

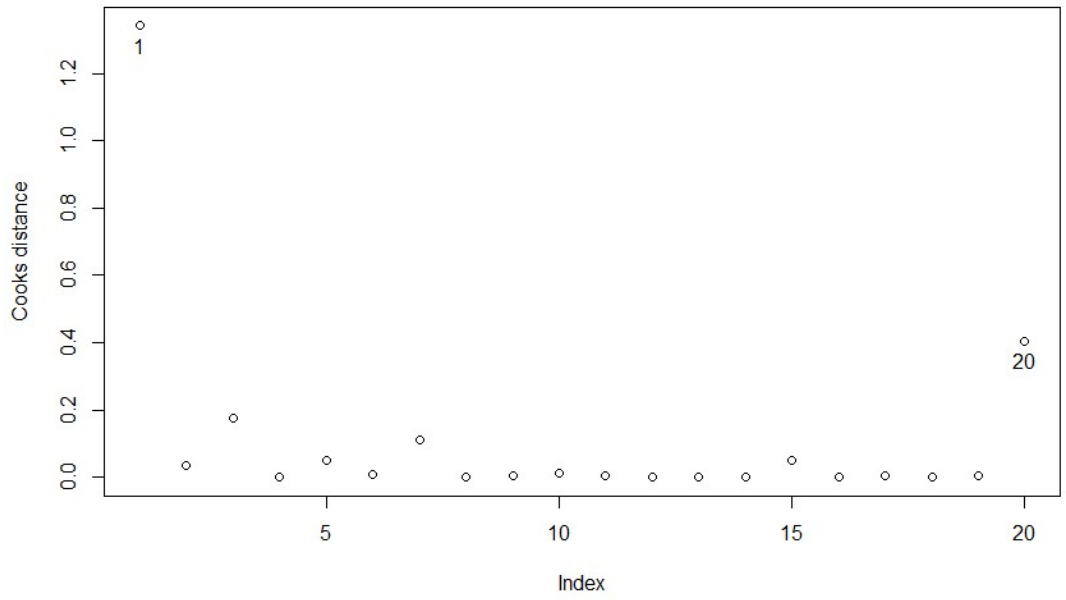
Leverage

```
> lev2[lev2 > 2*3/20]
        1         2         3
0.3261141 0.3435349 0.4479746
```

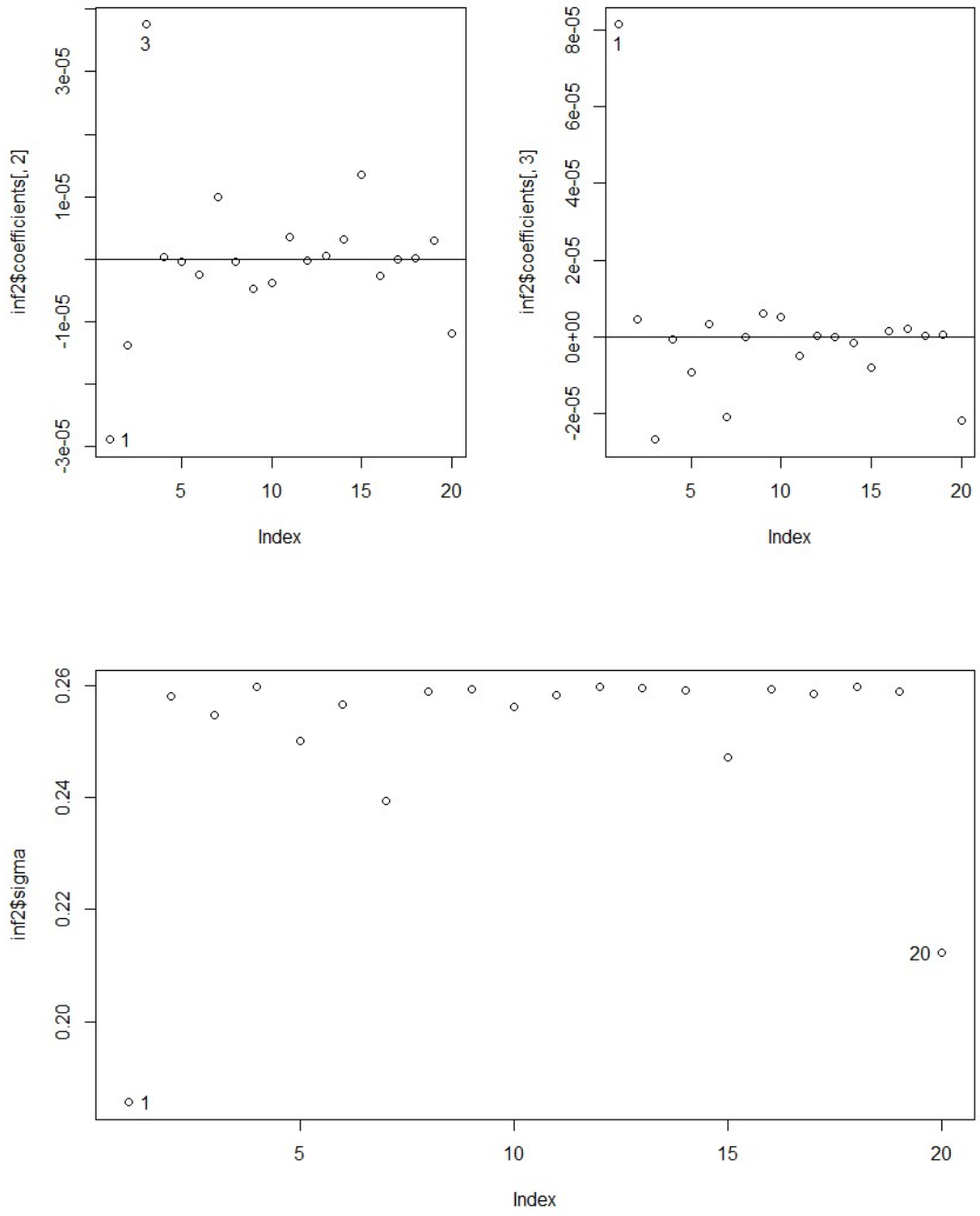
Studentized residual and Jacknife residual plot against index



Cook statistics against index



Influence of leave-out-one coefficients



For this model, we can more confidently state that observation with index 1 and 20 are outliers.

3.

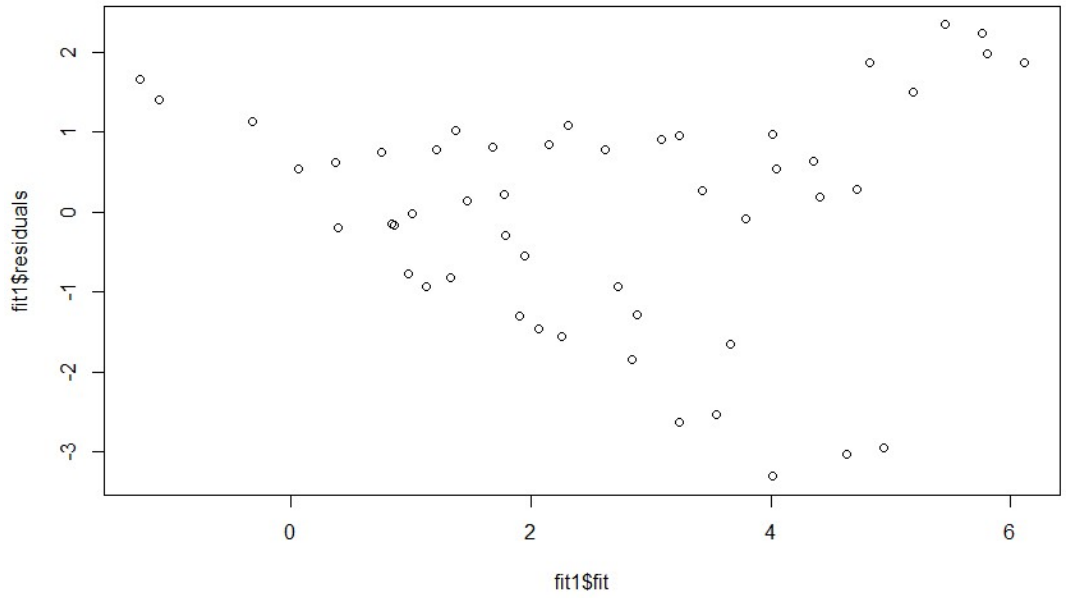
```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear
_models/hw6/acc.txt", header=T)
# ACC: acceleration of different vehicles.
```

```
# WHP: weight-to-horsepower ratio.
# SP: the speed at which they were traveling.
# G: the grade, G=0 implies the road was horizontal.
```

(a)

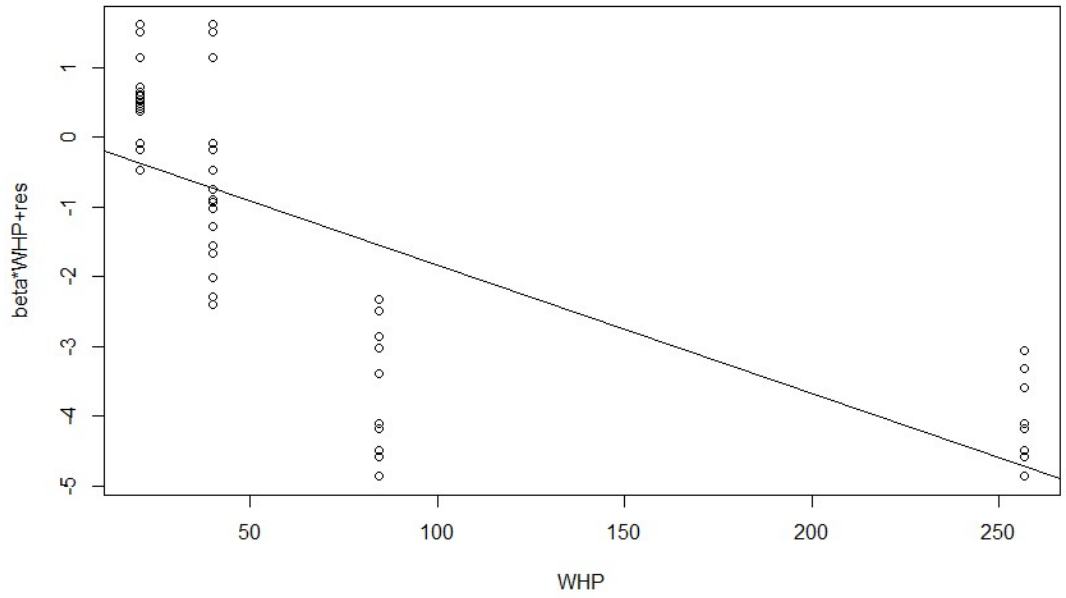
```
> fit1 <- lm(ACC ~ WHP + SP + G, data=gala)
> mean(fit1$residuals^2) # MSE
[1] 1.988278
```

Residual against ACC-hat

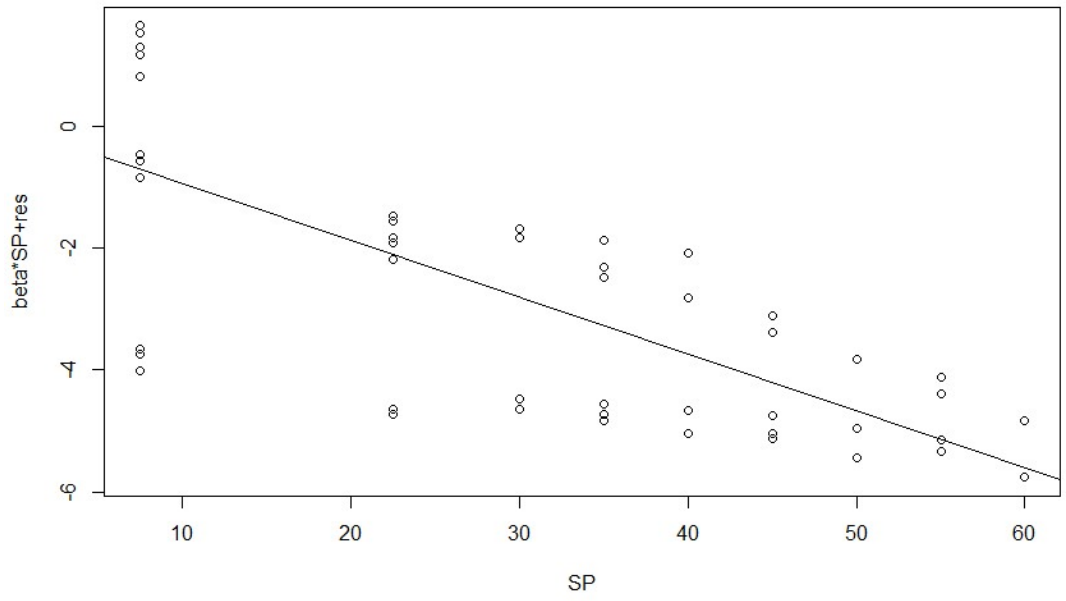


Partial residual plots

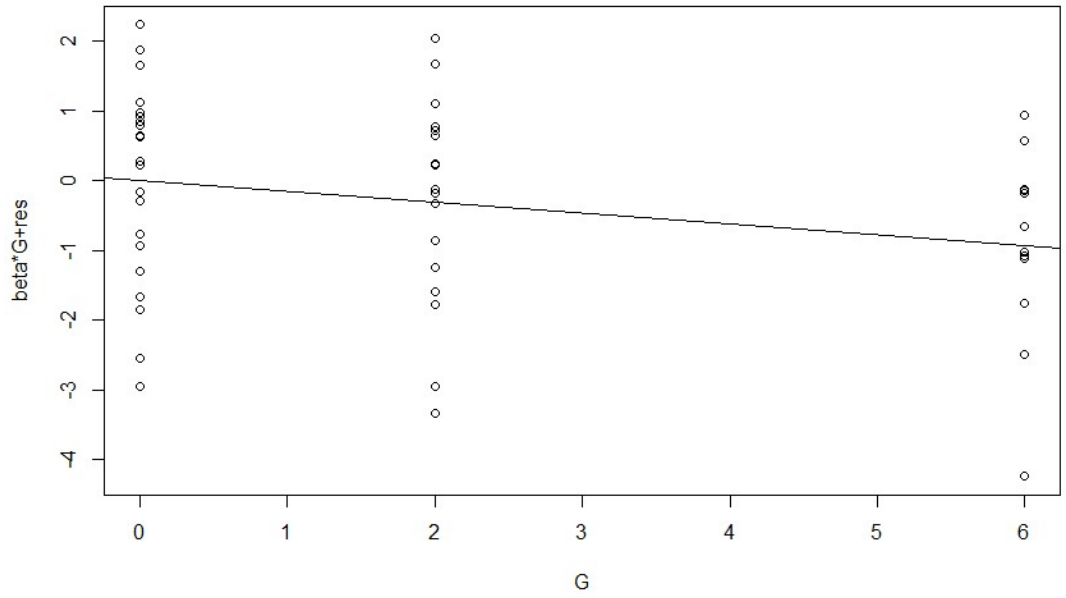
```
> prplot(fit1, 1)
```




```
> prplot(fit1, 2)
```




```
> prplot(fit1, 3)
```



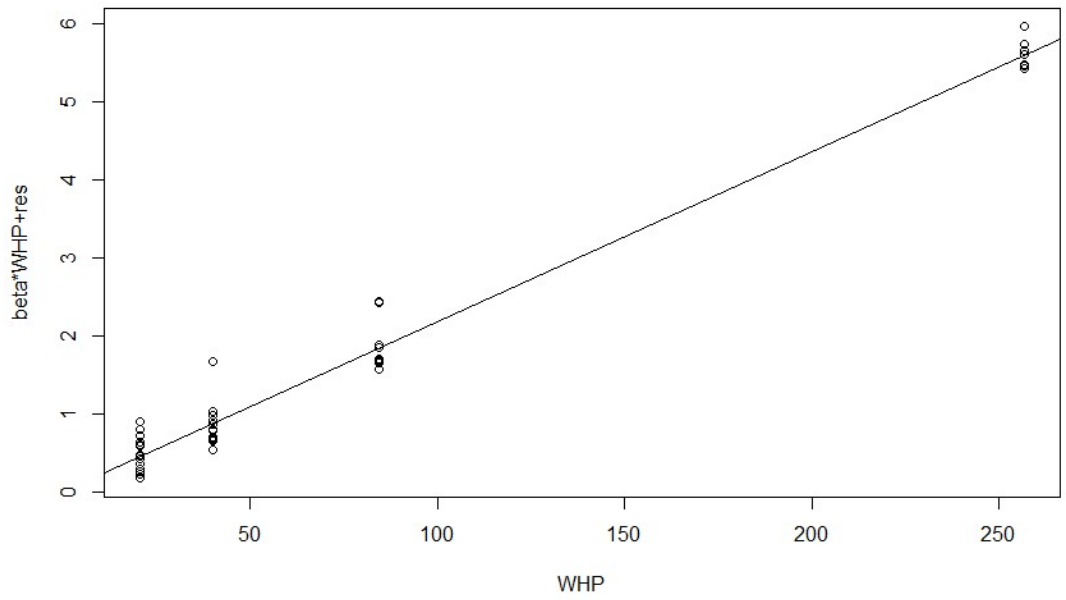
We can observe some kind of curvature of predictor “WHP”’s residual and a decreasing trend of variance for predictor “SP”. We might need to adjust our model to obey the constant variance assumption.

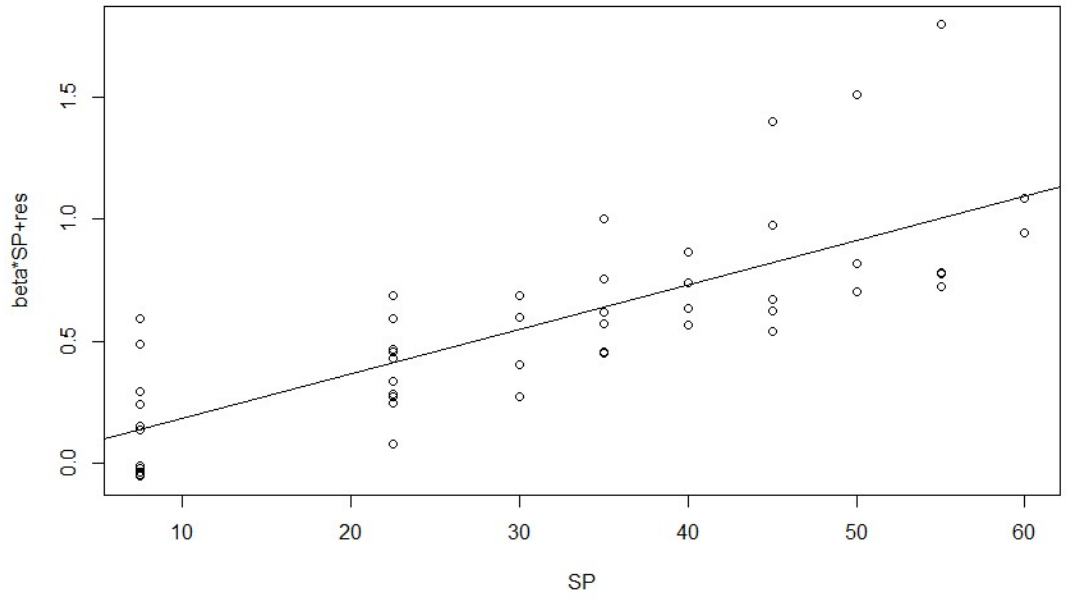
(b)

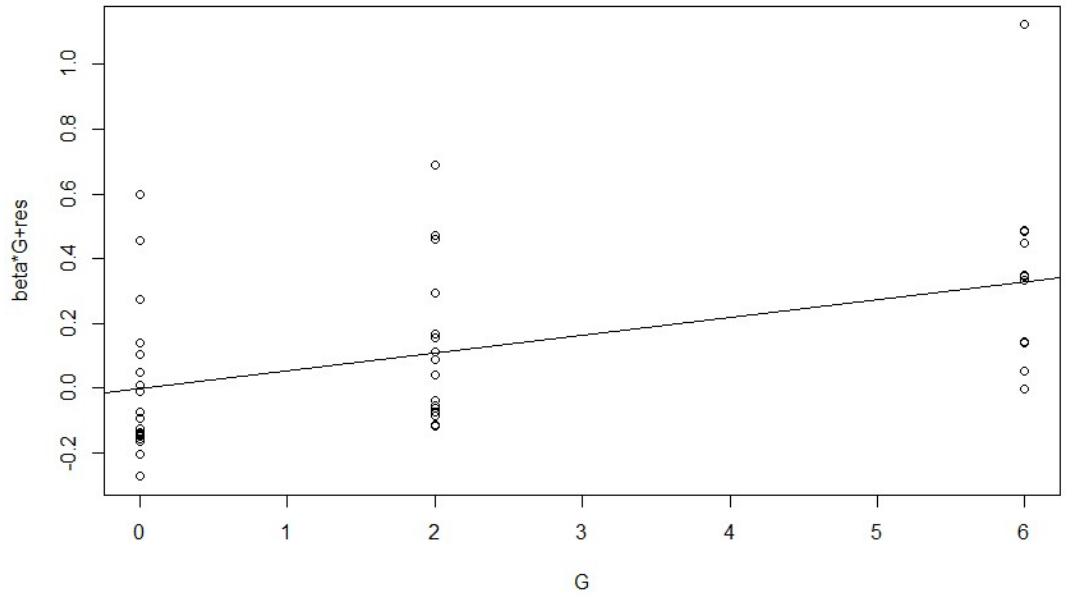
The best model as far as I can found is transforming the response by taking a square root and reciprocal, and adding a squared term of predictor “WHP”. It successfully eliminated the curvature in “WHP”’s partial residual and the variance of residual for “SP” is much uniform.

```
> ACCrec <- 1/ACC
> ACCrecsqrt <- ACCrec^0.5
> WHPsq <- WHP^2
> fitop <- lm(ACCrecsqrt ~ WHP + WHPsq + SP + G)
> mean(fitop$residuals^2) # MSE
[1] 0.05860664
```

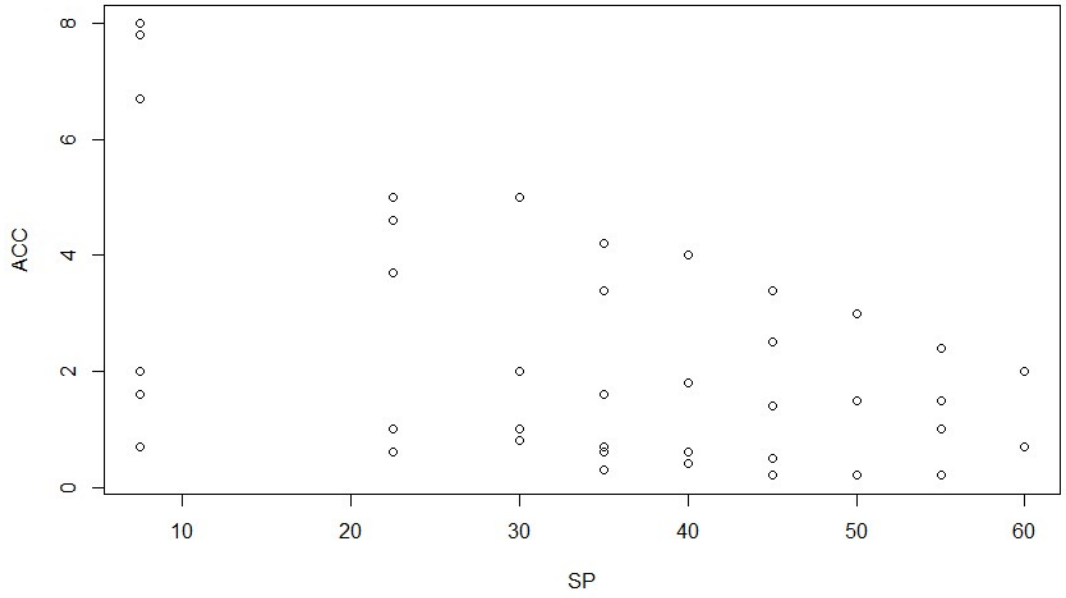
Partial residual plots


beta*WHP+res
0
1
2
3
4
5
6
50
100
150
200
250
WHP


beta*SP+res
0.0
0.5
1.0
1.5
10
20
30
40
50
60
SP



(c)



We can see that in the ACC plot against SP, there are lots of replicates. From the replicates we can roughly see the pure error is decreasing along with the mean. Therefore, when we fit a model, the residual will reveal the heteroscedasticity.