A2Q2

Undergraduate Student

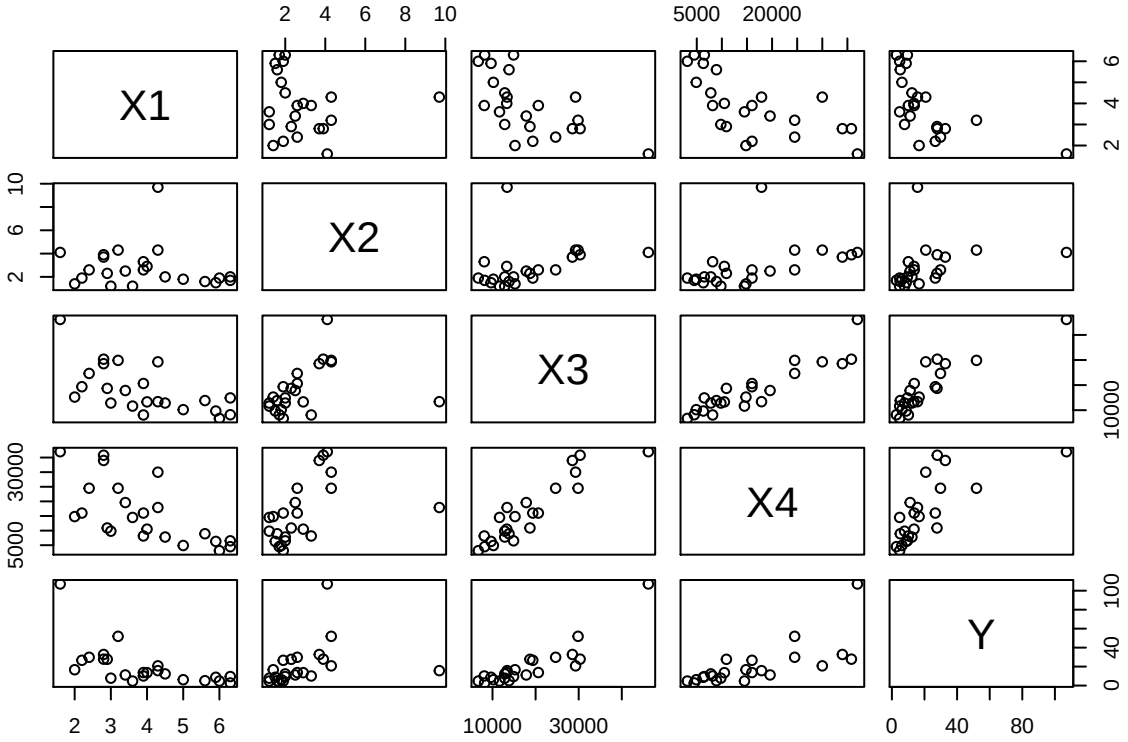
```
library(robustbase)
```

```
## Warning: package 'robustbase' was built under R version 4.0.5
```

```
data(aircraft)
```

(a)

```
pairs(aircraft)
```



Comments: From X2 column, we see there is one point that is far from other data points. This point can potentially be an outlier, same for X3 column and Y column. There is a possible collinearity between X3 and X4 because we see that data points tend to form an increasing line. Multilinearity is hard to detect from scatter plots.

(b)

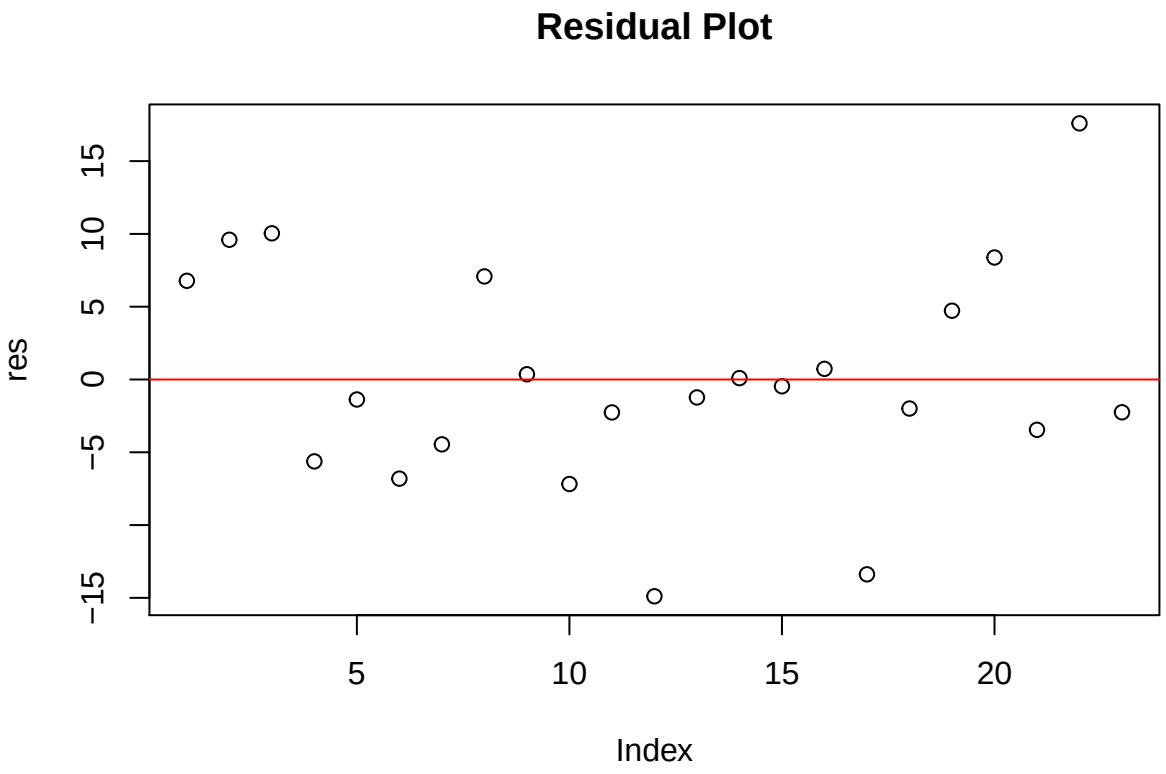
```
ols <- lm(Y~., data = aircraft)
summary(ols)
```

```
##
## Call:
## lm(formula = Y ~ ., data = aircraft)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.891  -3.955  -1.233   5.753  17.594
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.7913892 10.1157023  -0.375  0.71219
## X1          -3.8529189  1.7630016  -2.185  0.04232 *
## X2           2.4882665  1.1867538   2.097  0.05042 .
## X3           0.0034988  0.0004790   7.305 8.72e-07 ***
## X4          -0.0019537  0.0004986  -3.918  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.406 on 18 degrees of freedom
## Multiple R-squared:  0.8836, Adjusted R-squared:  0.8578
## F-statistic: 34.17 on 4 and 18 DF,  p-value: 3.501e-08
```

Comments: The ordinary least square regression model has a adjusted R-squared 0.8578, which is a fairly good number. This means that the model fits data well. We see that the p-value of X2 is 0.05042, which means that X2 is not statistically significant in the model, we may get rid of it.

(c)

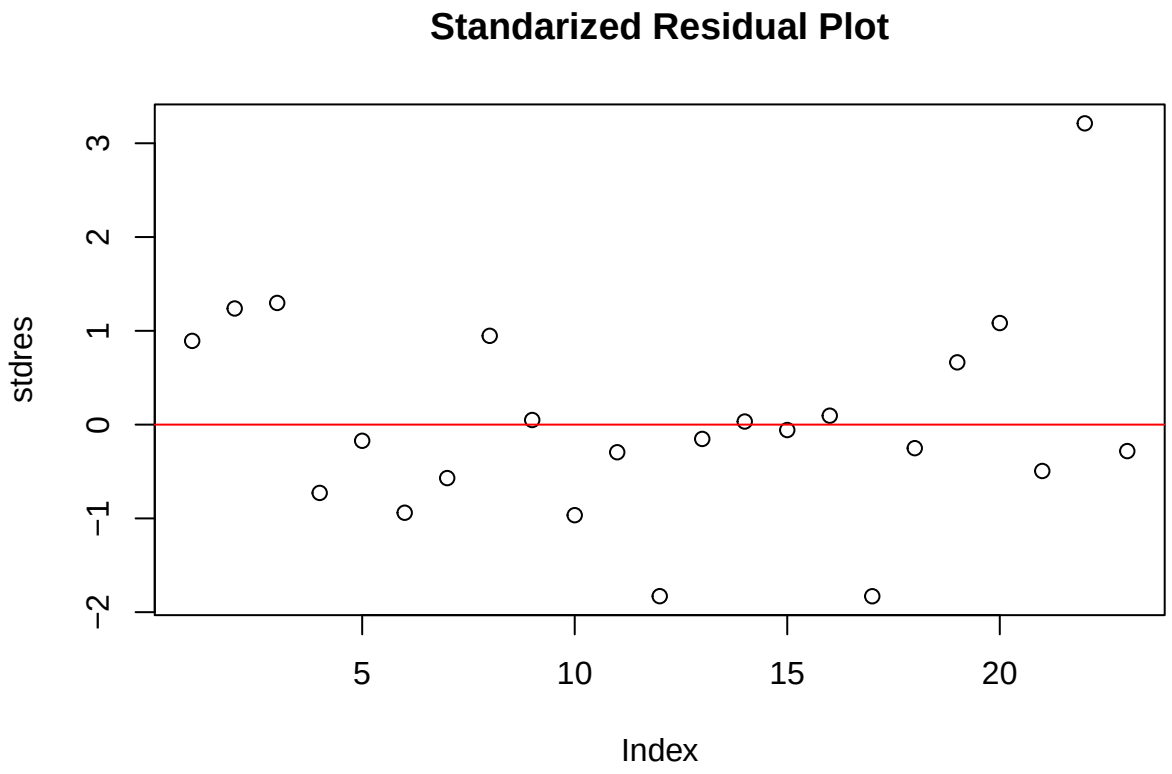
```
library(MASS)

res <- resid(ols)
plot(res, main = "Residual Plot")
abline(h=0, col='red')
```




```
stdres <- rstandard(ols)
# Point with large standarized residual
stdres[stdres>=2 | stdres<= -2]
```

```
##       22
## 3.212404
```

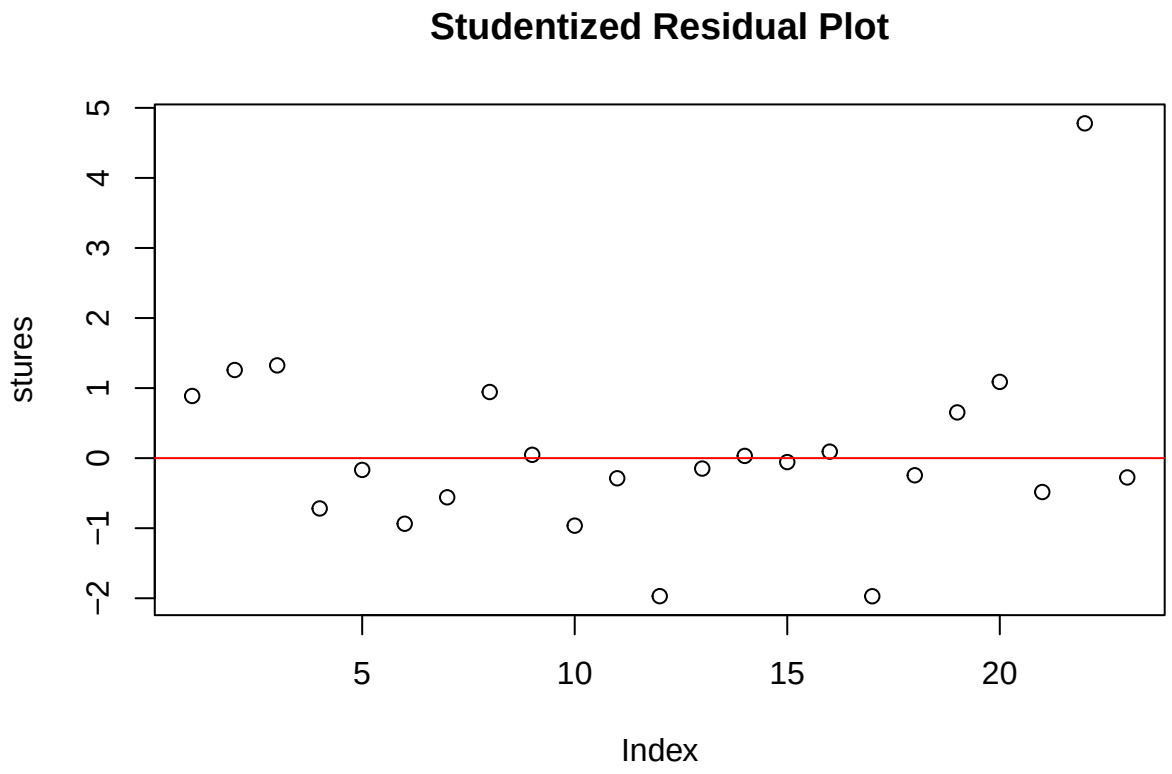
```
plot(stdres, main = "Standarized Residual Plot")
abline(h=0, col='red')
```




```
stures <- studres(ols)
# point with large studentized residual
stures[stures>=3 | stures<= -3]
```

```
##       22
## 4.779265
```

```
plot(stures, main = "Studentized Residual Plot")
abline(h=0, col='red')
```



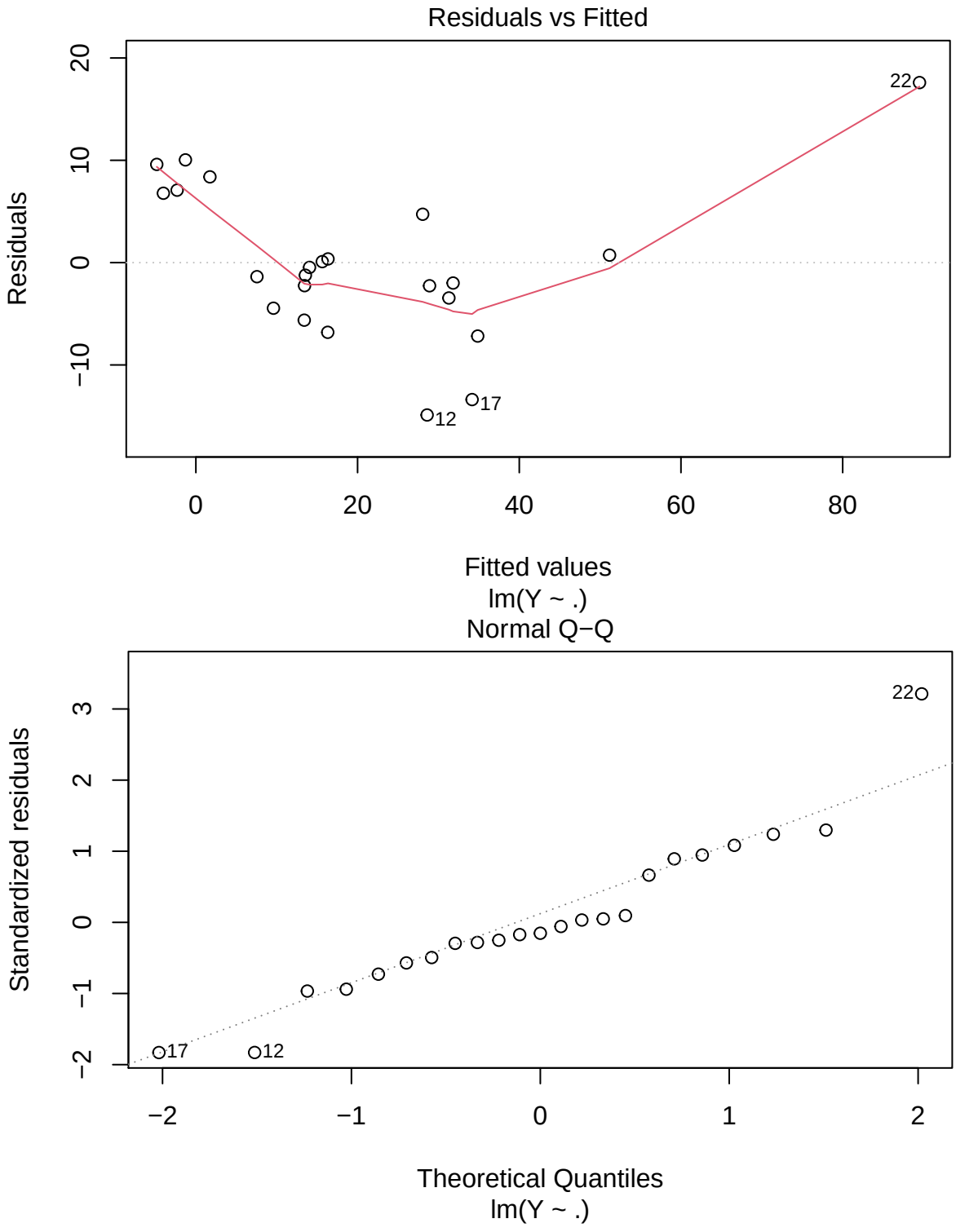

```
# Points with high leverage
leverage <- hatvalues(ols)
leverage[leverage>0.5]
```

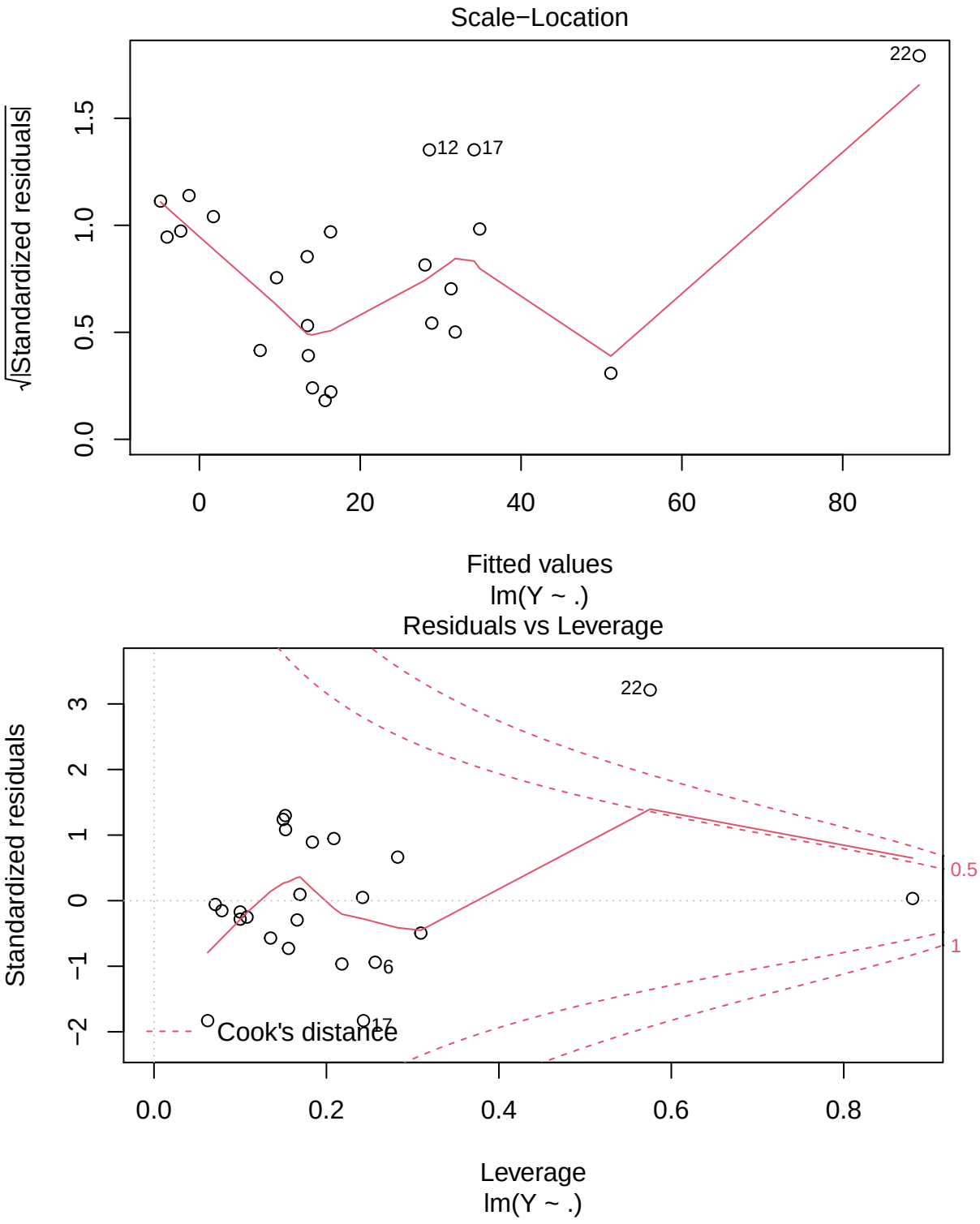
```
##        14        22
## 0.8798824 0.5754745
```

```
# Point with high cooks distance
cooks.distance(ols)[cooks.distance(ols)>=1]
```

```
##       22
## 2.797775
```

```
plot(ols)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Y ~ .)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ .)
22
17
12



Comments: From all three residuals we see that, point 22 has really high residual, and point 12, 17 have relatively high residual. From qq plot, we see that point 17 is actually on the line, but point 12 and point 22 are away from the line. For leverage calculation, we found that point 14 and point 22 have high leverage. For cooks distance calculation, we found that point 22 have distance greater than 1. Therefore, we can conclude that point 22 is a leverage outlier with high influential. Point 14 is a leverage point but is non-influential. Point 17 and point 12 are outliers and have relatively high residual but is in reasonable range.

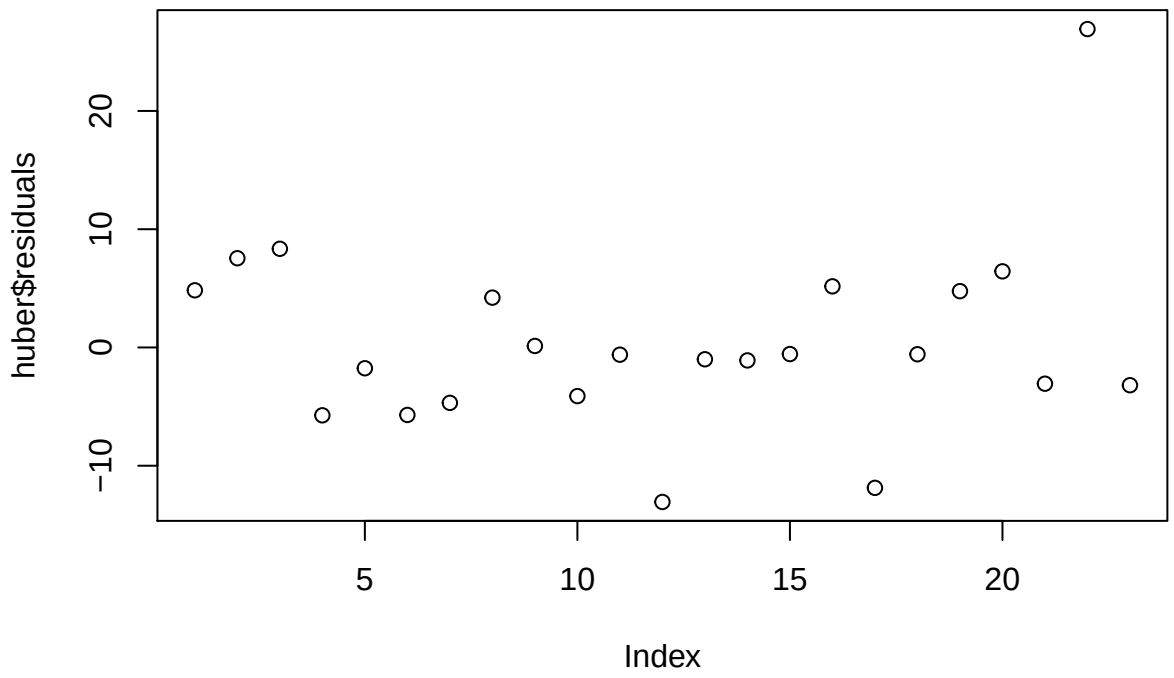
(d)

(i)

```
huber <- rlm(Y~., data = aircraft, maxit=50)
summary(huber)
```

```
##
## Call: rlm(formula = Y ~ ., data = aircraft, maxit = 50)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0636  -3.6520  -0.6103   4.7975  26.9243
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) -1.2850  8.6035    -0.1494
## X1          -3.4214  1.4994    -2.2818
## X2           2.2160  1.0093     2.1955
## X3           0.0029  0.0004     7.2207
## X4          -0.0016  0.0004    -3.6940
##
## Residual standard error: 6.946 on 18 degrees of freedom
```

```
plot(huber$residuals)
```



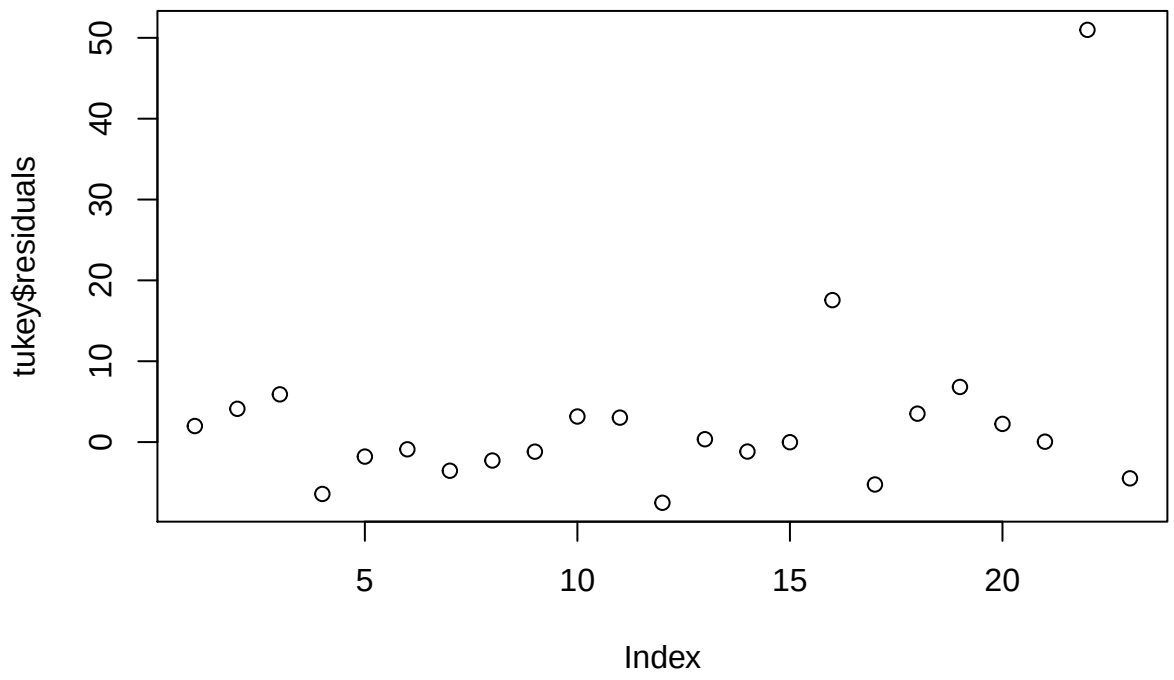

```
huber$residuals[huber$residuals> 20 | huber$residuals< -20]
```

```
##      22
## 26.9243
```

```
tukey <- rlm(Y~., psi = psi.bisquare, maxit=50, data = aircraft)
summary(tukey)
```

```
##
## Call: rlm(formula = Y ~ ., data = aircraft, psi = psi.bisquare, maxit = 50)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.50246 -2.03569  0.04858  3.34014 50.98828
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  8.7164  6.5184     1.3372
## X1          -3.1804  1.1361    -2.7995
## X2           1.3792  0.7647     1.8035
## X3           0.0016  0.0003     5.1288
## X4          -0.0007  0.0003    -2.2088
##
## Residual standard error: 4.689 on 18 degrees of freedom
```

```
plot(tukey$residuals)
```




```
tukey$residuals[tukey$residuals> 20 | tukey$residuals< -20]
```

```
##       22
## 50.98828
```

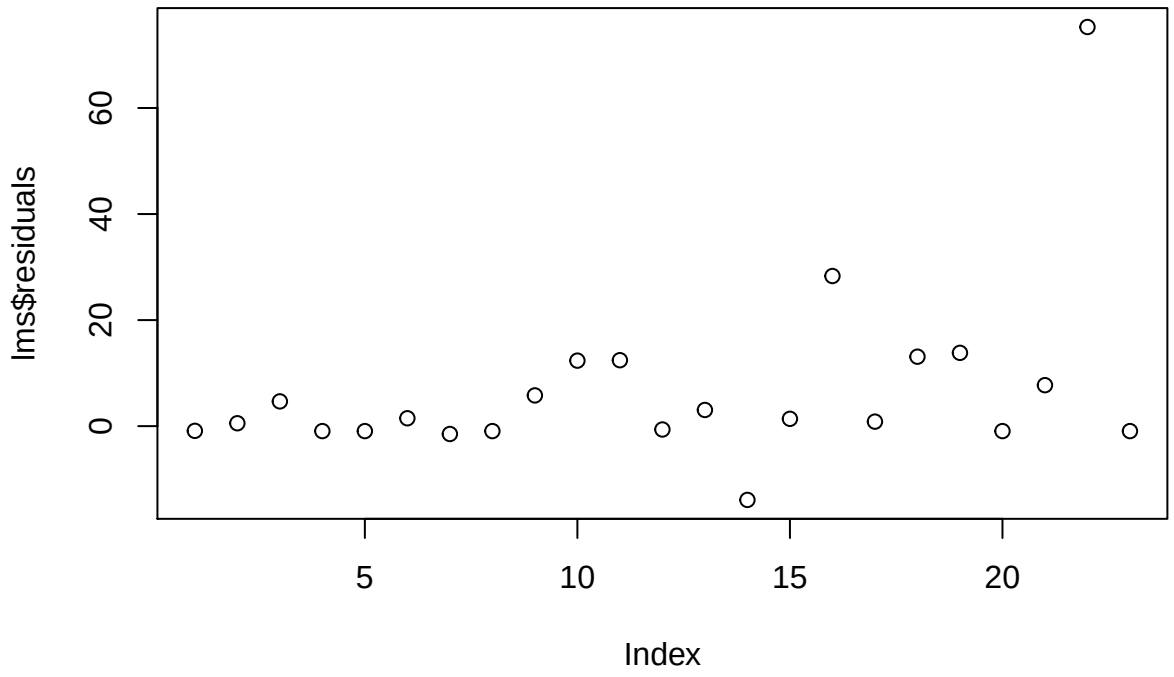
Comments: Both method reduce residual standard error drastically, while tukey has lower residual standard error than huber. For huber, point 22 is an outlier and influential, and it is an influential point. For tukey, point is are outlier and influential.

(ii)

```
lms <- lqs(Y~., data = aircraft, method = "lms")
lms
```

```
## Call:
## lqs.formula(formula = Y ~ ., data = aircraft, method = "lms")
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##   4.9876707   -1.5847221    2.9371076    0.0006037   -0.0002849
##
## Scale estimates 2.093 2.211
```

```
plot(lms$residuals)
```




```
lms$residuals[lms$residuals> 20 | lms$residuals< -20]
```

```
##       16       22
## 28.31029 75.27270
```

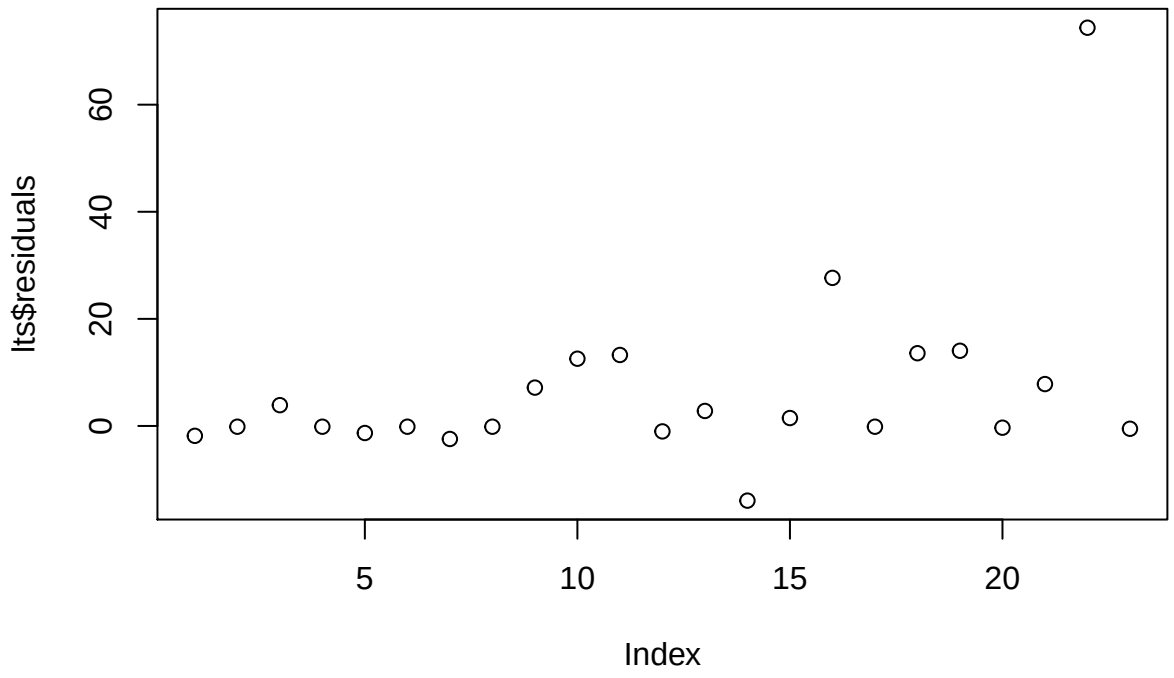
Comments: The coefficient of X2 becomes really influential to the prediction, most of the residuals are around 0. The model Point 16, 22 are outliers. Point 16 is non-influential and point 22 is influential.

(iv)

```
lts <- lqs(Y~., data = aircraft)
lts
```

```
## Call:
## lqs.formula(formula = Y ~ ., data = aircraft)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##   1.3620702   -0.9800261    2.9808583    0.0007157   -0.0003333
##
## Scale estimates 3.485 3.695
```

```
plot(lts$residuals)
```




```
lts$residuals[lts$residuals> 20 | lts$residuals< -20]
```

```
##       16       22
## 27.65469 74.37083
```

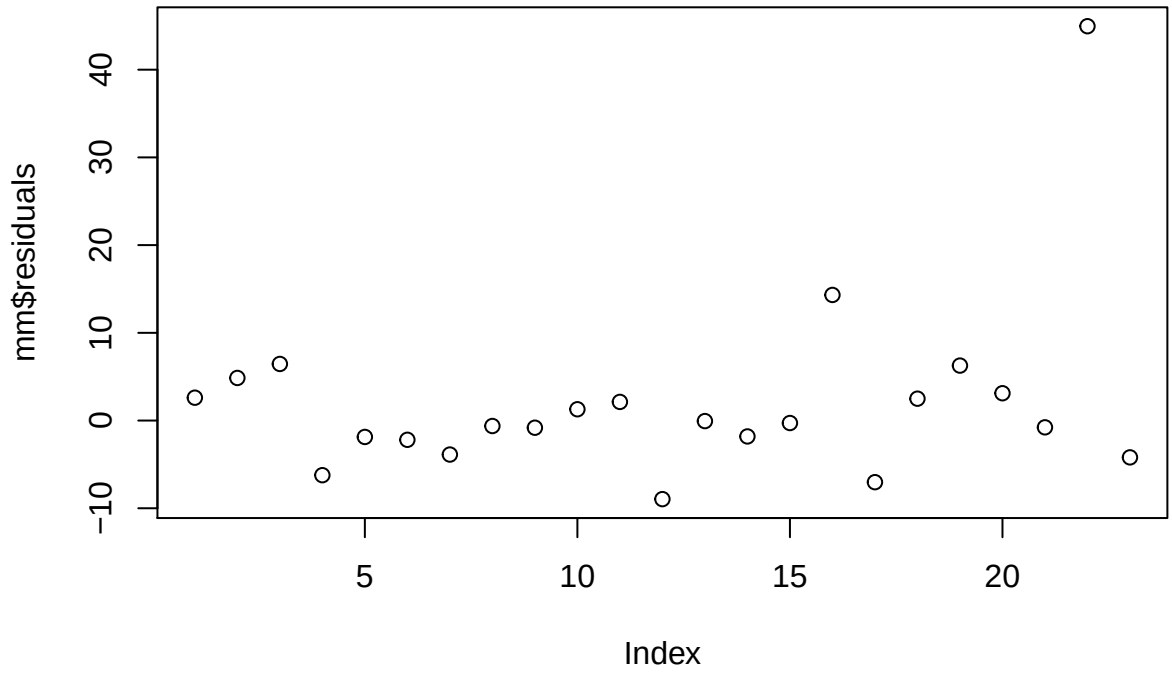
Comments: The coefficient of X2 is reduced comparing to lms, more absolute value of residuals are within 10. The only outlier is point 22 and it is influential.

(v)

```
mm <- rlm(Y~., data = aircraft, method = "MM")
summary(mm)
```

```
##
## Call: rlm(formula = Y ~ ., data = aircraft, method = "MM")
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9584 -2.0323 -0.2836  2.8585 44.9534
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  6.1417  6.8385     0.8981
## X1          -3.2306  1.1918    -2.7106
## X2           1.6711  0.8023     2.0830
## X3           0.0019  0.0003     5.9329
## X4          -0.0009  0.0003    -2.7562
##
## Residual standard error: 5.892 on 18 degrees of freedom
```

```
plot(mm$residuals)
```




```
mm$residuals[mm$residuals> 20 | mm$residuals< -20]
```

```
##       22
## 44.95345
```

Comments: MM model combines lms and lts, the coefficient of X1 becomes dominant. we found that the residual standard error is between huber and tukey. The value of intercept is drastically reduced comparing to lms and lts. Point 22 is an outlier and influential. The residual of point 22 is lower than lms and lts.

(vi)

```
library(quantreg)
```

```
## Warning: package 'quantreg' was built under R version 4.0.5
```

```
## Loading required package: SparseM
```

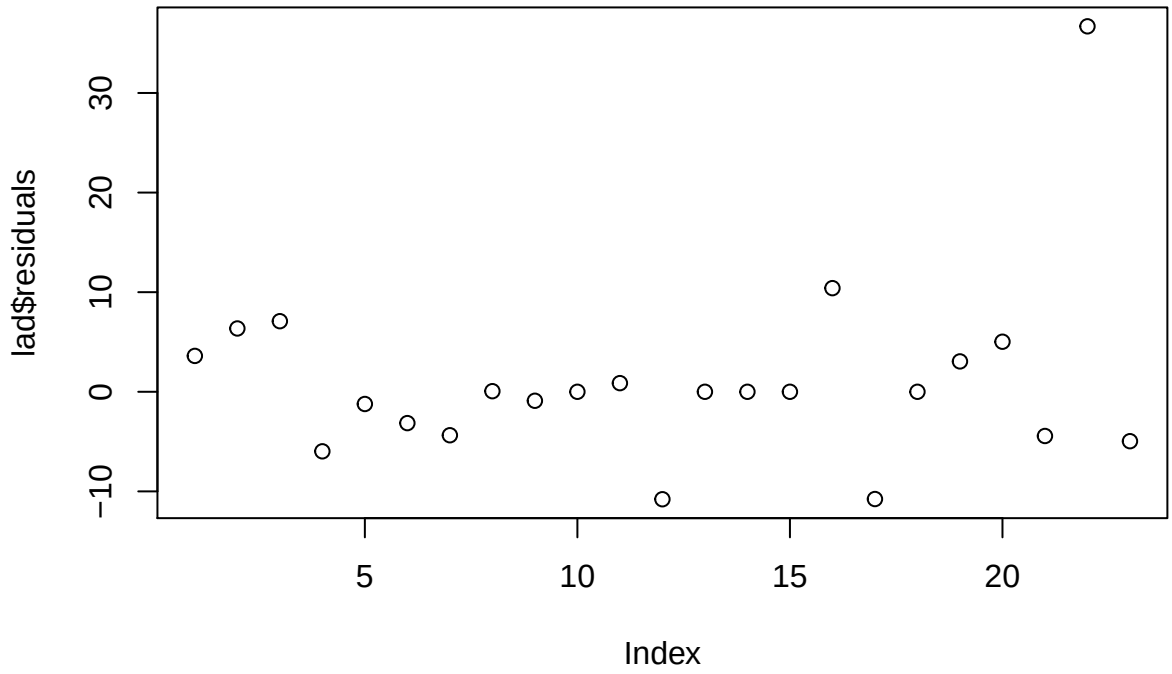
```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
lad <- rq(Y~., data = aircraft, tau = 0.5)
summary(lad)
```

```
##
## Call: rq(formula = Y ~ ., tau = 0.5, data = aircraft)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept)   1.95110    -28.72698  12.14364
## X1           -3.04663     -5.87179   0.38236
## X2            1.47359    -12.01175   7.39919
## X3            0.00223      0.00099   0.00395
## X4           -0.00096     -0.00237  -0.00056
```

```
plot(lad$residuals)
```



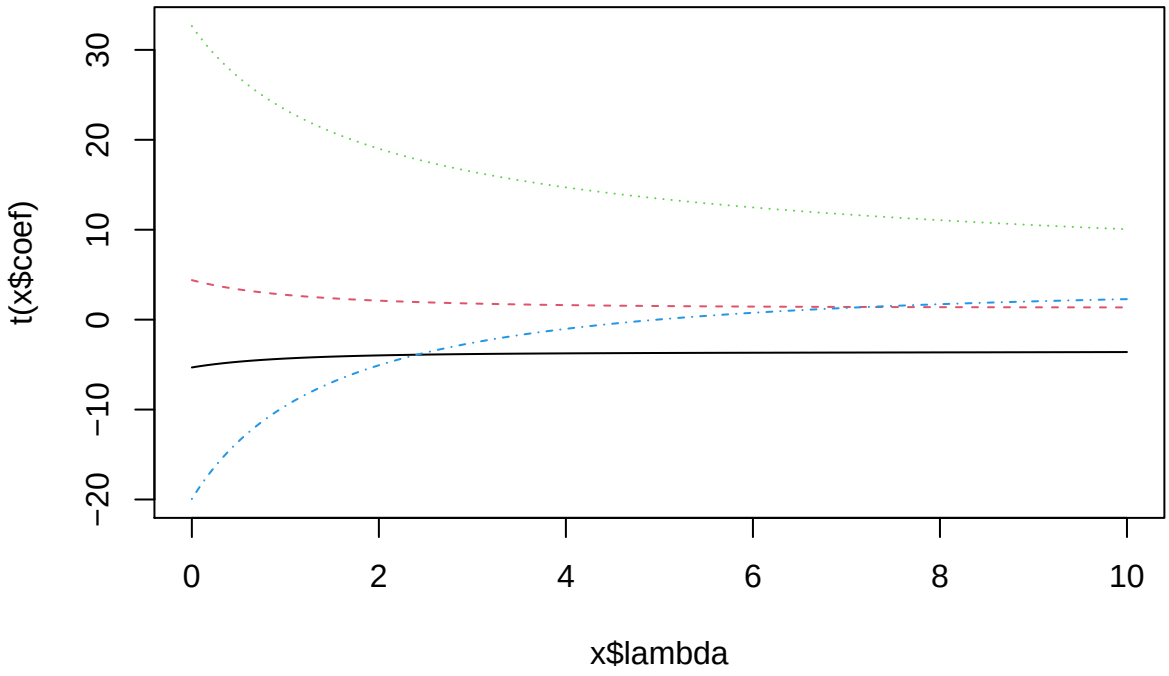
```
lad$residuals[lad$residuals> 15 | lad$residuals< -15]
```

```
##       22
## 36.68868
```

Comment: We see that the coefficient of X1 still have the most influence. Nearly all absolute value of residuals are within 10. Point 22 is an outlier and it is an influential point.

(e)

```
ridge <- lm.ridge(Y~., data = aircraft, lambda = seq(0, 10, 0.001))
plot(ridge)
```




```
select(ridge)
```

```
## modified HKB estimator is 0.09351538
## modified L-W estimator is 0.3365202
## smallest value of GCV  at 0.067
```

```
ridge.fit <- lm.ridge(Y~., data = aircraft,lambda = c(0.09351538, 0.3365202, 0.067))
ols$coefficients
```

```
##  (Intercept)           X1           X2           X3           X4
## -3.791389154 -3.852918860  2.488266504  0.003498787 -0.001953669
```

```
ridge.fit
```

```
##                              X1       X2          X3           X4
## 0.09351538 -3.619233 -3.742094 2.350837 0.003358387 -0.001807903
## 0.33652020 -3.185760 -3.511161 2.059600 0.003054408 -0.001495020
## 0.06700000 -3.667744 -3.772073 2.388142 0.003396665 -0.001847573
```

Comments: We observe large fluctuation for small K, even for the smallest value of GCV. This means that the collinearity (or possibly multicollinearity) exists and has a great impact on the result of OLS.