hw5

Hoang Chu

2023-10-11

```
library(faraway)
```

6.1

```
data("sat")
summary(sat)
```

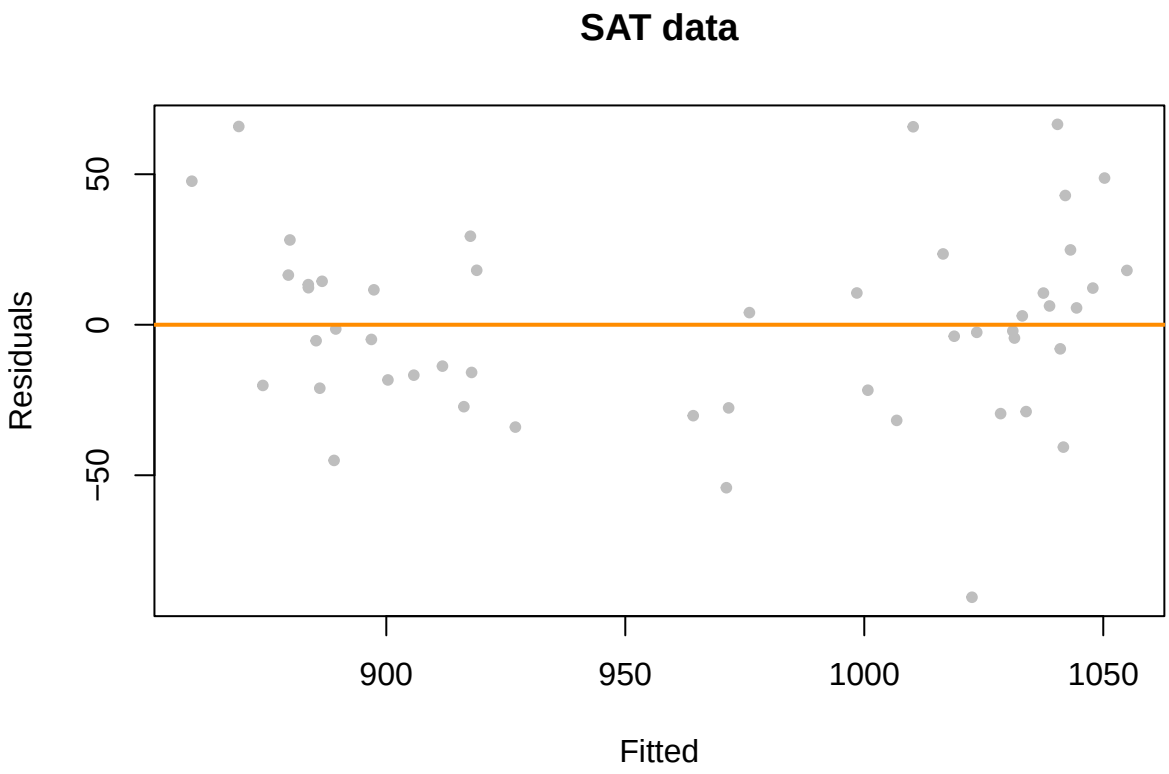
```
##      expend          ratio           salary          takers
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
##      verbal           math           total
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :448.0   Median :497.5   Median : 945.5
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

```
m <- lm(total ~ expend + ratio + salary + takers, data=sat)
```

a.

```
residuals <- resid(m)
```

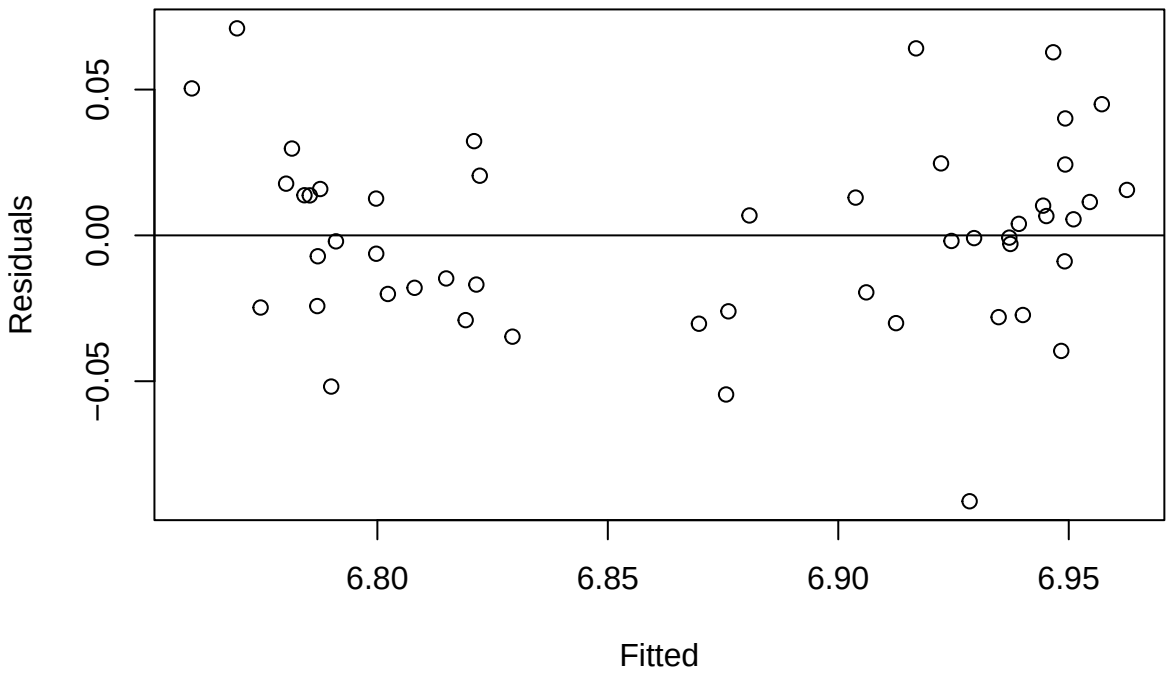
```
plot(fitted(m), residuals, col="grey", pch=20, xlab="Fitted", ylab="Residuals", main = "SAT data")
abline(h=0, col="darkorange", lwd=2)
```



From the residual-fitted plot, spread of residuals is roughly constant across the range of fitted values, hence the equal variance assumption is satisfied. However, the plot indicates some non-linearity.

We can improve non-linearity by taking the log.

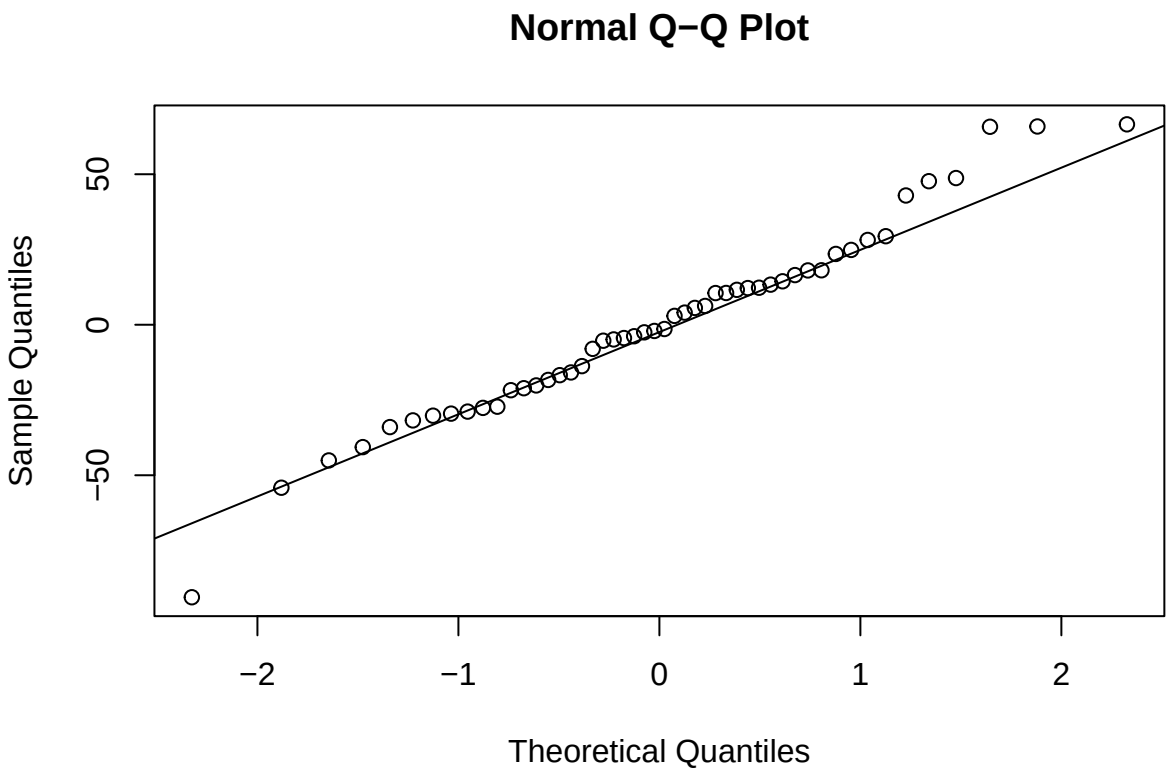
```
log_m <- lm(log(total) ~ expend + salary + ratio + takers, data = sat)
plot(fitted(log_m), residuals(log_m), xlab = "Fitted", ylab = "Residuals")
abline(h=0)
```



As variances hover more and more around 0, the non-linearity has been improved.

b.

```
qqnorm(residuals)
qqline(residuals)
```




```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.97691, p-value = 0.4304
```

Since the shapiro p-value is more than 0.05, we fail the reject the null hypothesis of the test which is the residuals are normally distributed. Hence the normally assumption is satisfied.

c.

```
hatvalues(m)[hatvalues(m) > 2 * mean(hatvalues(m))]
```

```
##  California Connecticut  New Jersey        Utah
##   0.2821179   0.2254519   0.2220978   0.2921128
```

We can see that [California, Connecticut, New Jersey, and Utah] are leverage points.

d.

```r
rstandard(m)[abs(rstandard(m)) > 2]
```

```
## New Hampshire  North Dakota          Utah West Virginia
##      2.103095      2.123567      2.390264     -2.858505
```

We can see that [New Hampshire, North Dakota, Utah, and West Virginia] are outliers.

Since Utah appeared in both leverage points and outliers, it has a significant pull to our linear model. Hence, it's better to remove it.

e.

```r
cook_dist <- cooks.distance(m)
```

```r
cook_dist['New Hampshire'] > 4 / length(cook_dist)
```

```
## New Hampshire
##         FALSE
```

```r
cook_dist['North Dakota'] > 4 / length(cook_dist)
```

```
## North Dakota
##        FALSE
```

```r
cook_dist['Utah'] > 4 / length(cook_dist)
```

```
## Utah
## TRUE
```

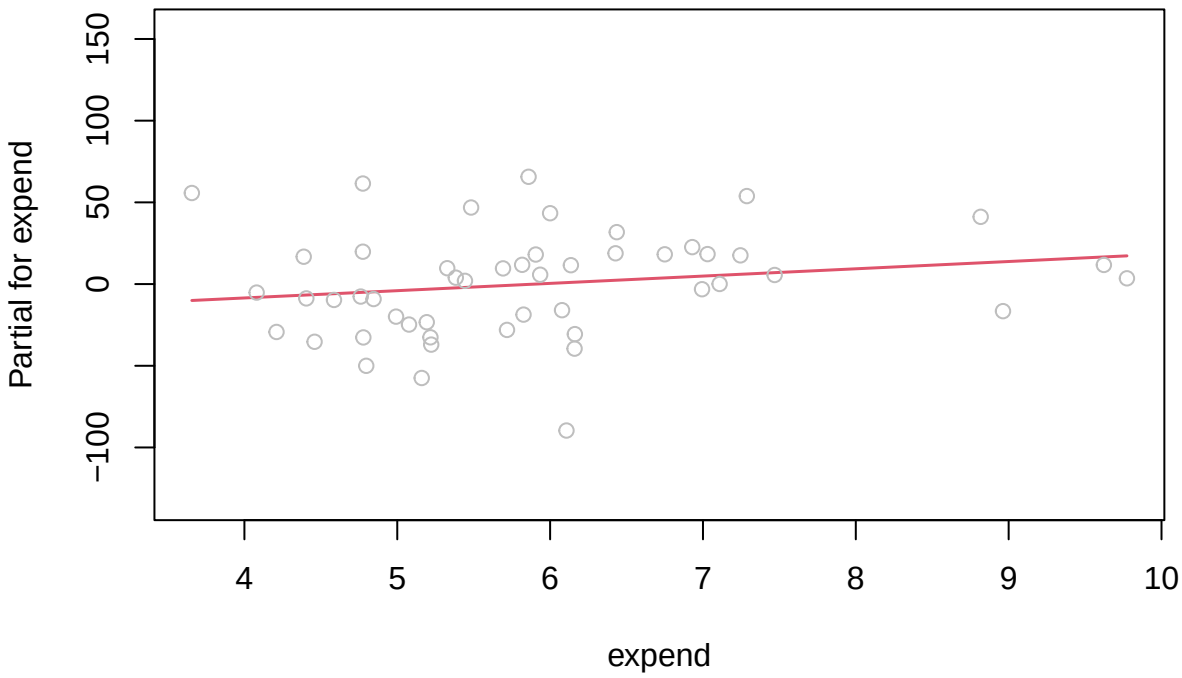
```r
cook_dist['West Virginia'] > 4 / length(cook_dist)
```

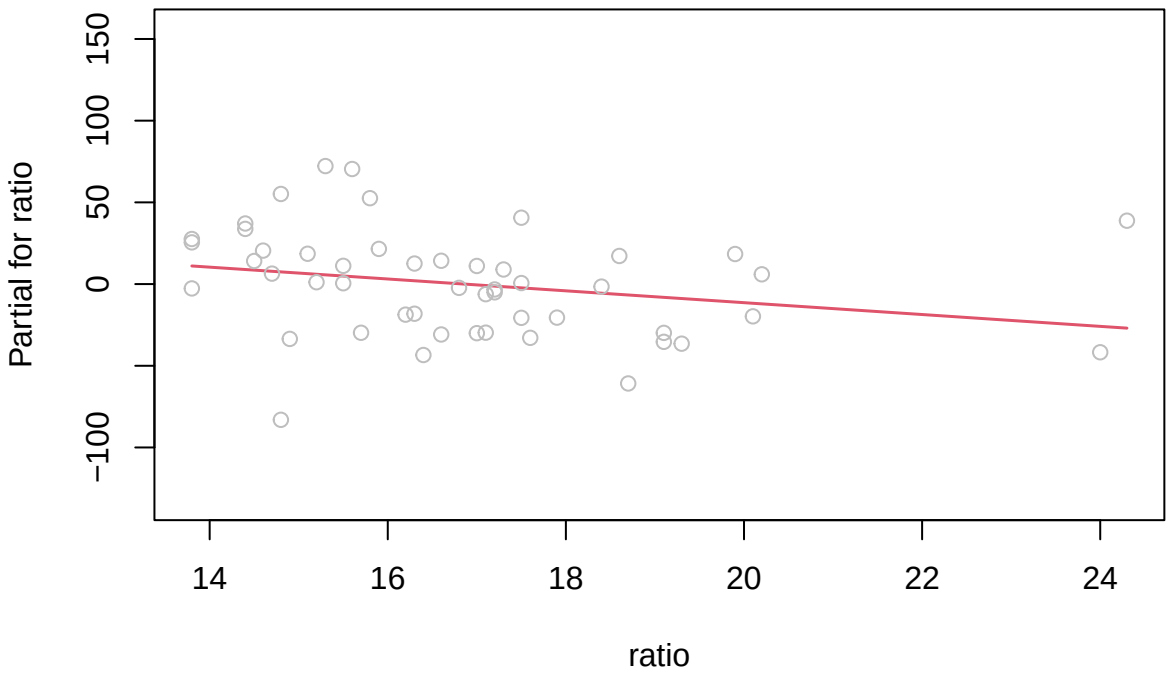
```
## West Virginia
##          TRUE
```

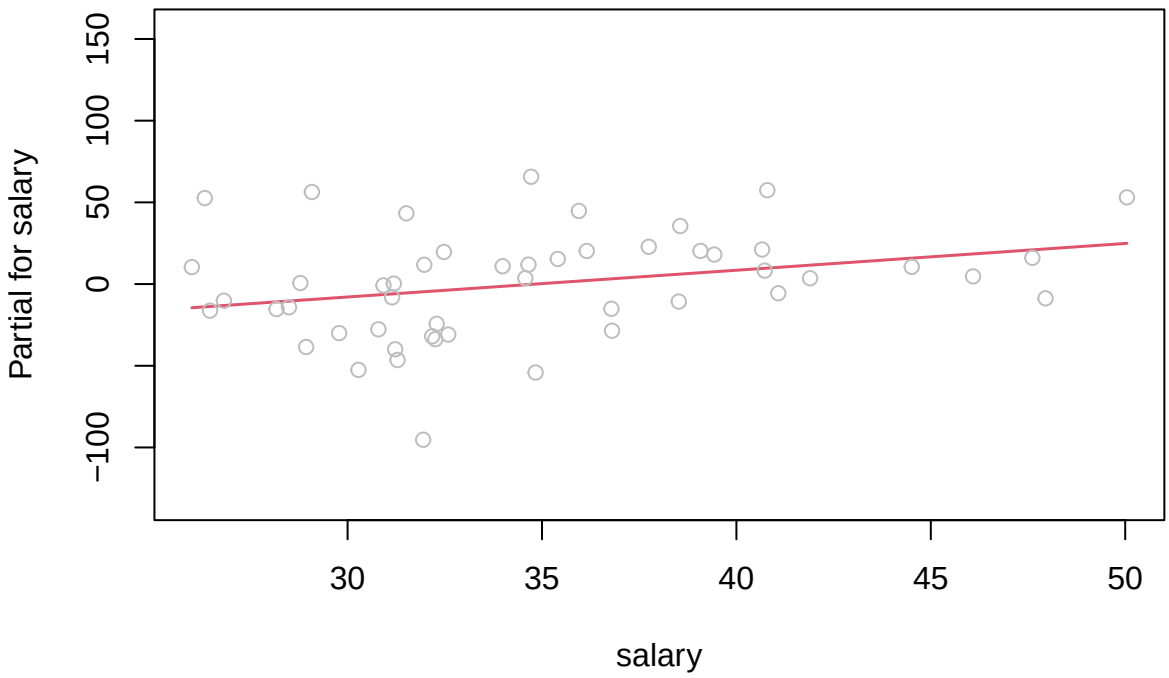
We can see that only [Utah, West Virginia] are influential points, and should be removed.

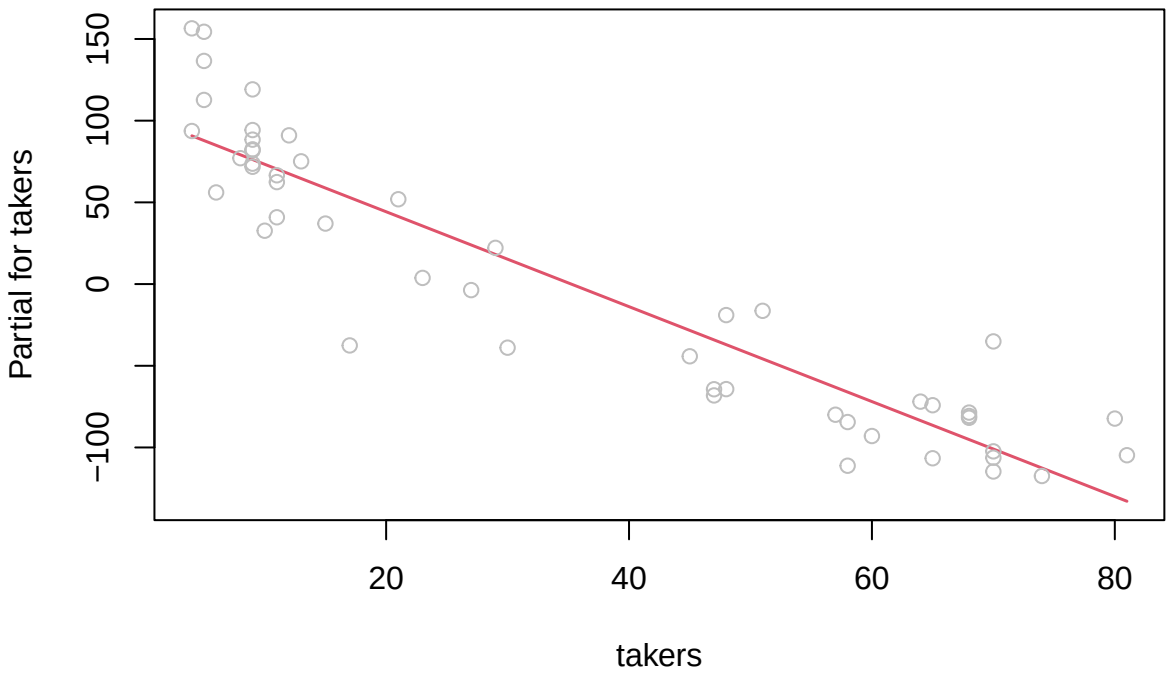
f.

```r
termplot(m,partial.resid = T)
```


Partial for expend
150
100
50
0
−100
4
5
6
7
8
9
10
expend


Partial for ratio
150
100
50
0
−100
14
16
18
20
22
24
ratio


Partial for salary
150
100
50
0
−100
30
35
40
45
50
salary



The flatness of the lines associated with three of the variables (expend, salary, and ratio) reflect their lack of significance.

The model may be improved by dropping them.

6.3

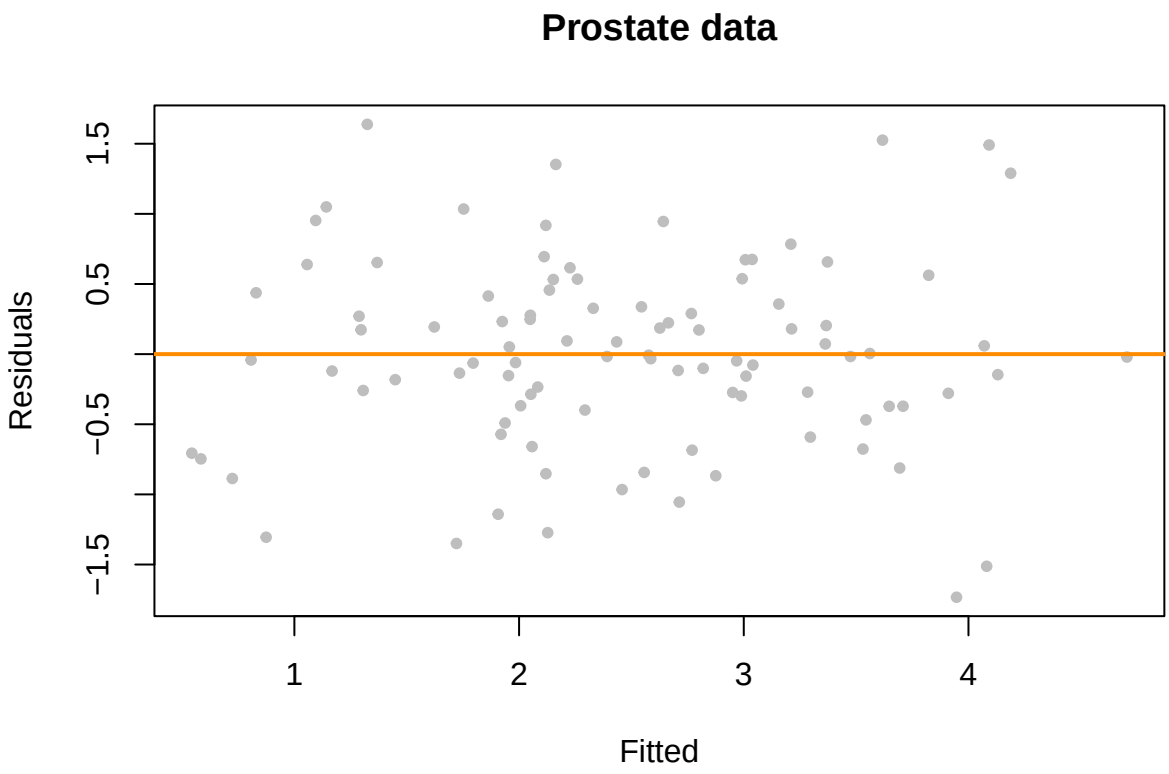
```
data("prostate")
```

```
m <- lm(lpsa ~ ., data=prostate)
```

a.

```
residuals <- resid(m)
```

```
plot(fitted(m), residuals, col="grey", pch=20, xlab="Fitted", ylab="Residuals", main = "Prostate data")
abline(h=0, col="darkorange", lwd=2)
```

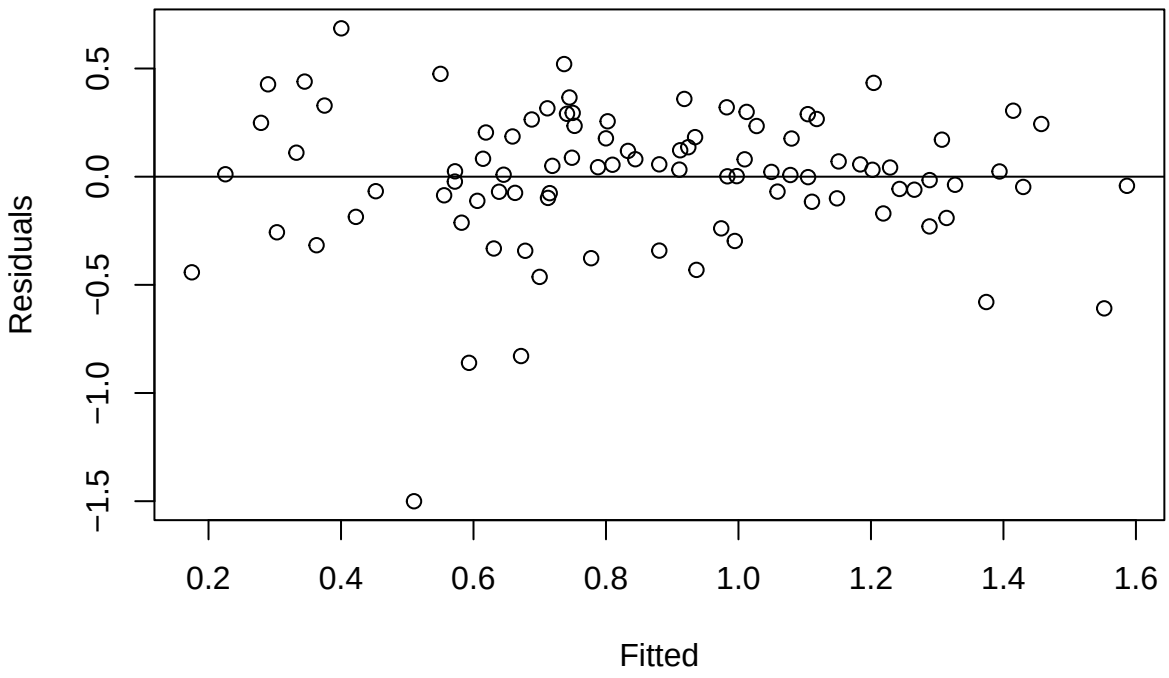


From the residual-fitted plot, spread of residuals is roughly constant across the range of fitted values, hence the equal variance assumption is satisfied, and the linear assumption is also satisfied as variances hover around 0.

```
log_m <- lm(log(lpsa) ~ ., data=prostate)
```

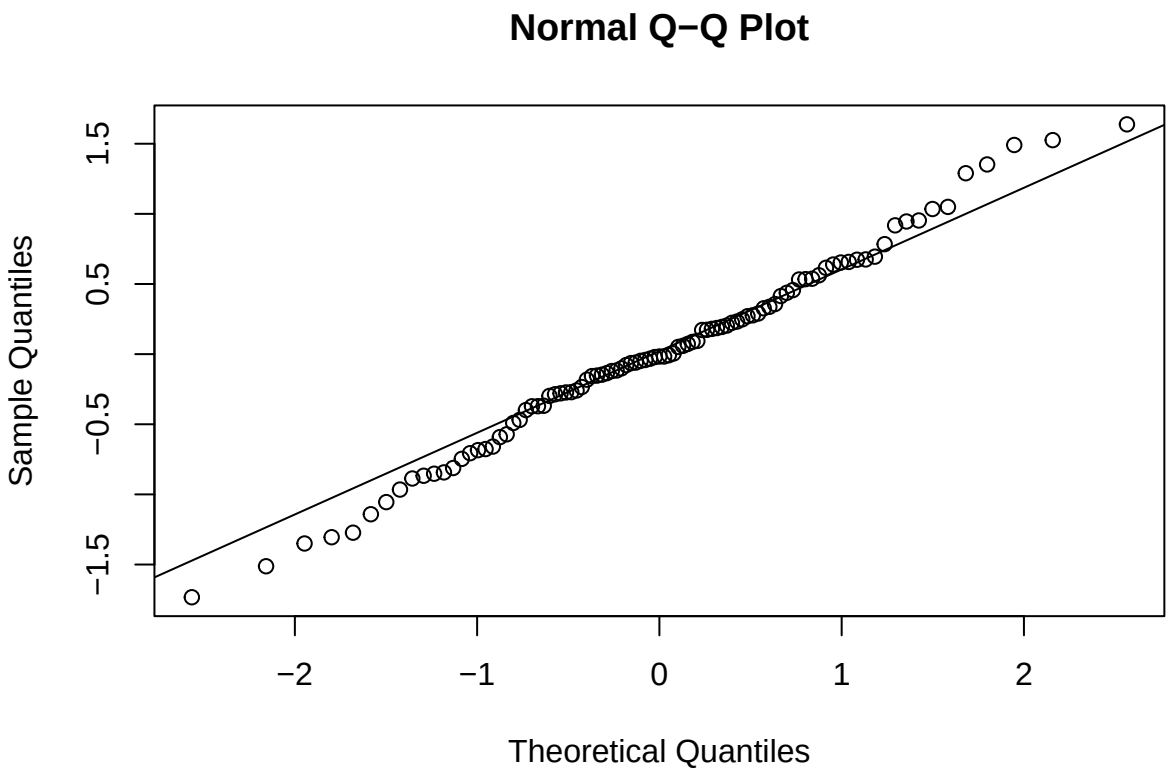
```
## Warning in log(lpsa): NaNs produced
```

```
plot(fitted(log_m), residuals(log_m), xlab = "Fitted", ylab = "Residuals")
abline(h=0)
```



b.

```
qqnorm(residuals)
qqline(residuals)
```




```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.99113, p-value = 0.7721
```

Since the shapiro p-value is more than 0.05, we fail the reject the null hypothesis of the test which is the residuals are normally distributed. Hence the normally assumption is satisfied.

c.

```
hatvalues(m)[hatvalues(m) > 2 * mean(hatvalues(m))]
```

```
##        32        37        41        74        92
## 0.3304757 0.2184392 0.2410079 0.1912109 0.2092421
```

We can see that observations [32, 37, 41, 74, 92] are leverage points.

d.

```r
rstandard(m)[abs(rstandard(m)) > 2]
```

```
##        39        47        69        95        97
## -2.534124 -2.316280  2.477016  2.323964  2.239719
```

We can see that [39, 47, 69, 95, 97] are outliers.

e.

```r
cook_dist <- cooks.distance(m)
```

```r
cook_dist[39] > 4 / length(cook_dist)
```

```
##   39
## TRUE
```

```r
cook_dist[47] > 4 / length(cook_dist)
```

```
##   47
## TRUE
```

```r
cook_dist[69] > 4 / length(cook_dist)
```

```
##   69
## TRUE
```

```r
cook_dist[95] > 4 / length(cook_dist)
```

```
##   95
## TRUE
```

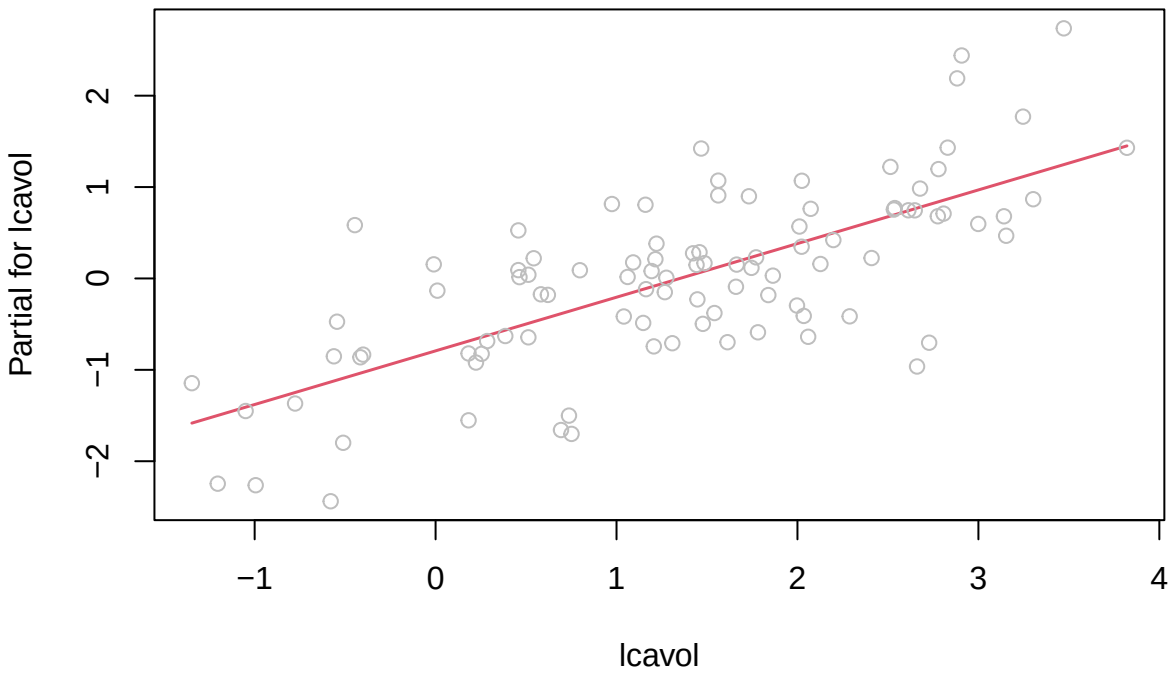
```r
cook_dist[97] > 4 / length(cook_dist)
```

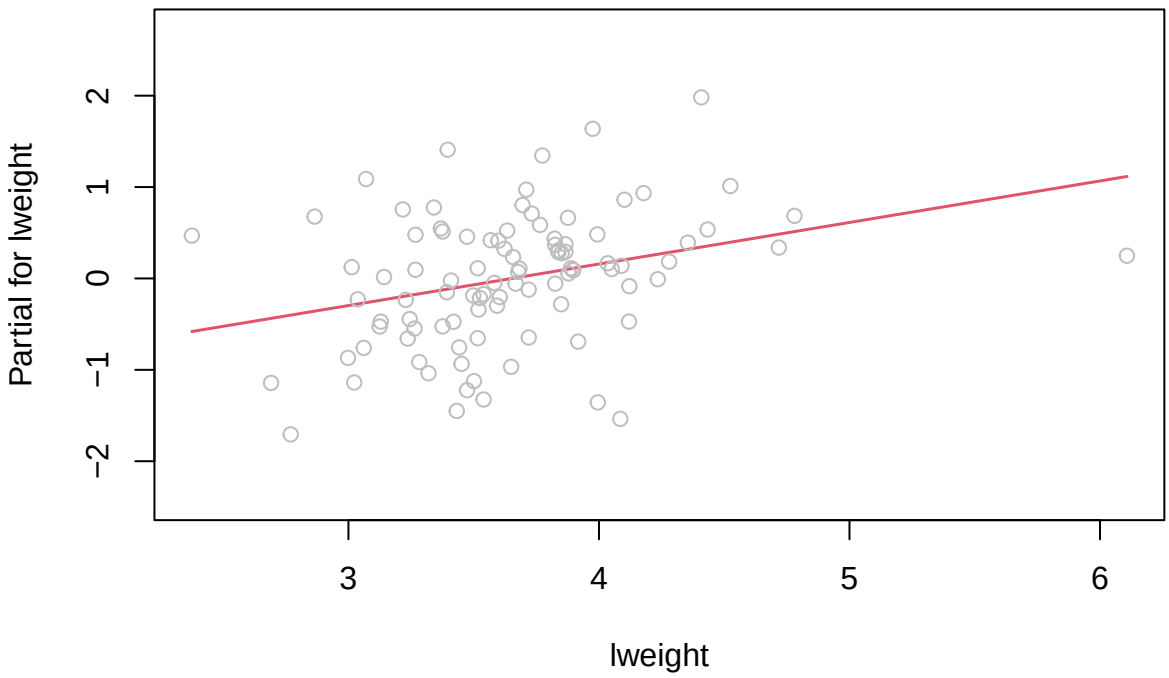
```
##   97
## TRUE
```

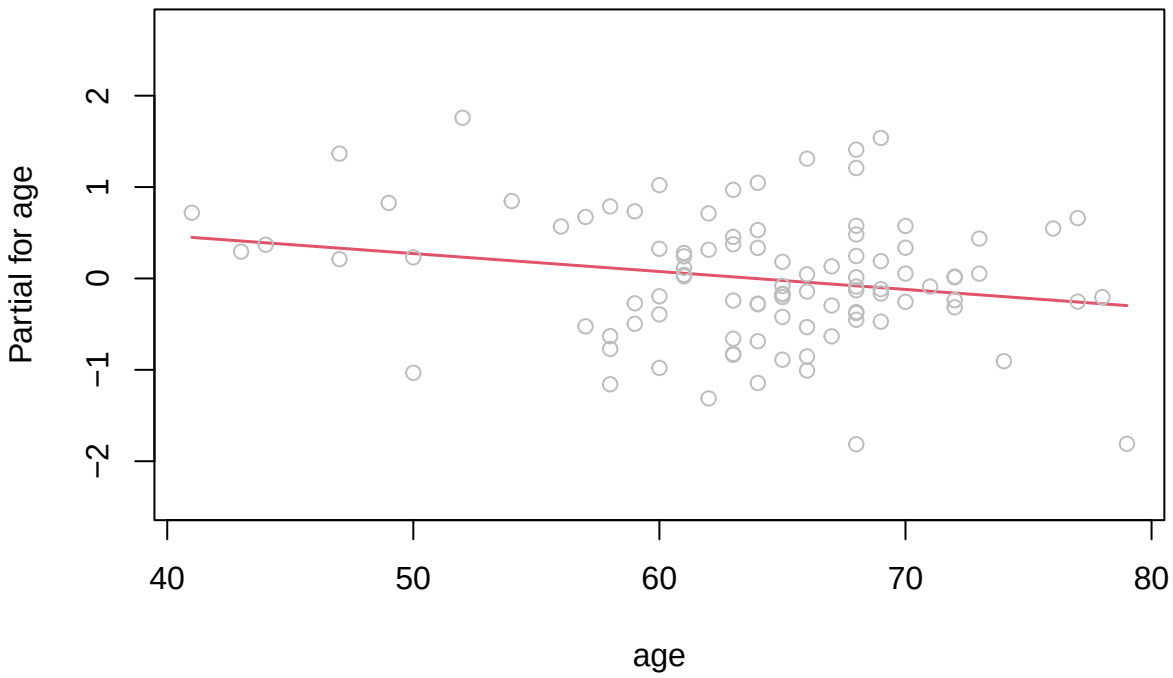
We can see that all those outliers are influential points, hence we should drop them.

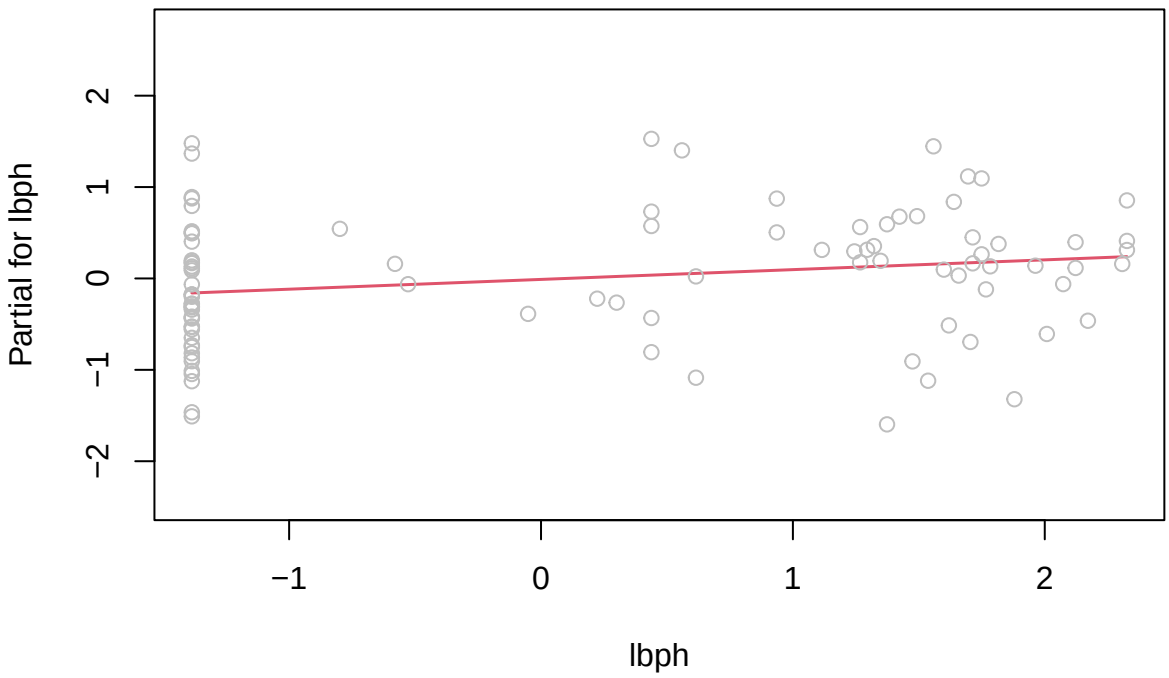
f.

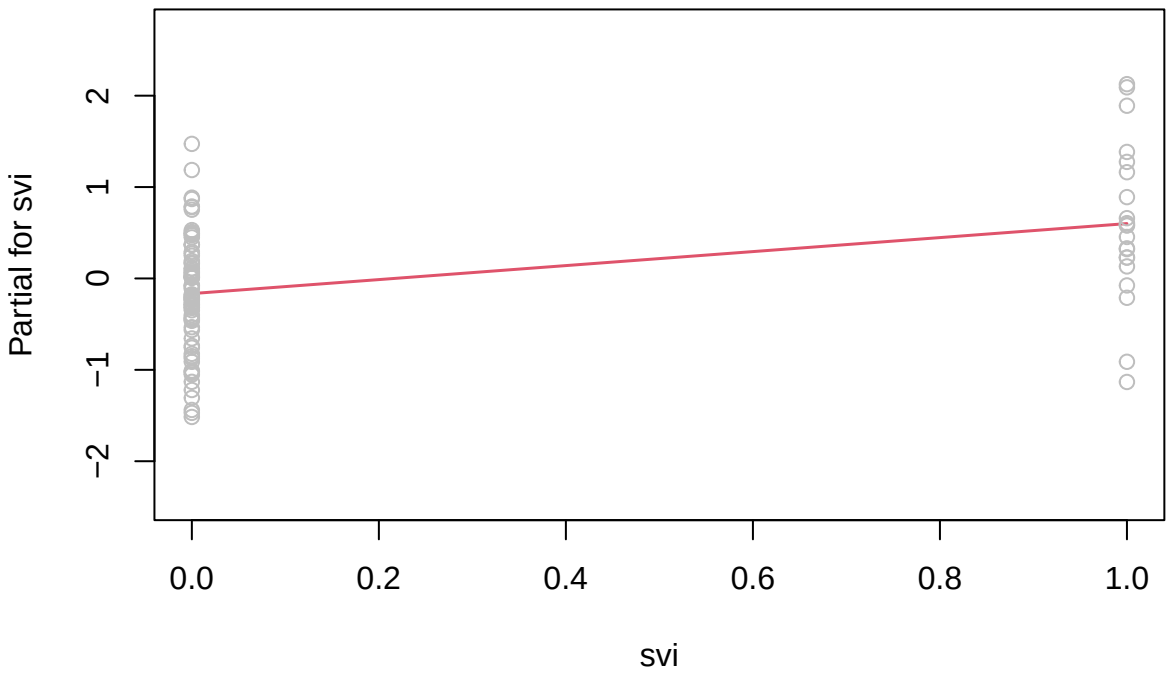
```r
termplot(m,partial.resid = T)
```

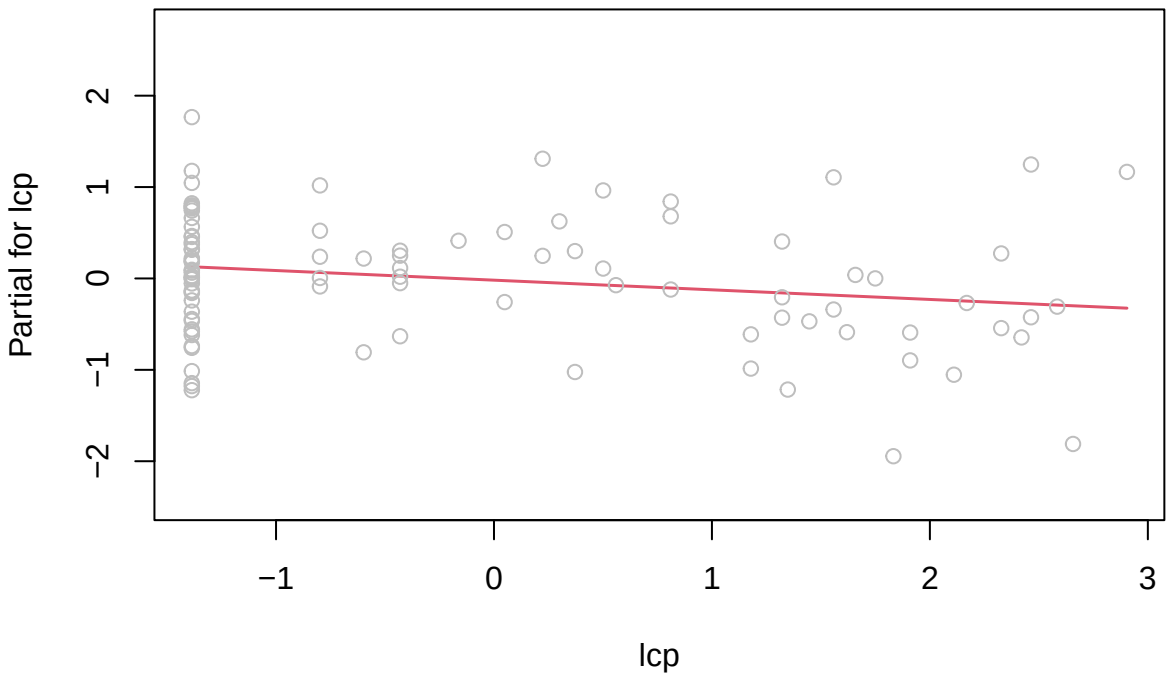

Partial for lcavol
2
1
0
−1
−2
−1
0
1
2
3
4
lcavol

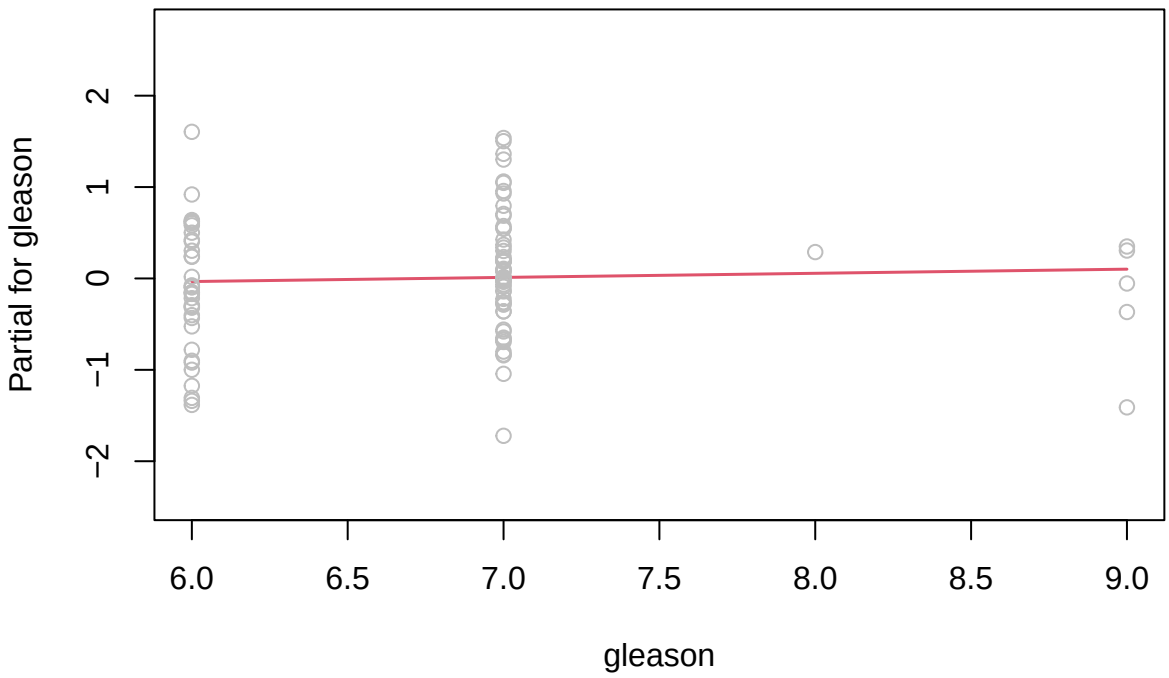

Partial for lweight
2
1
0
−1
−2
3
4
5
6
lweight

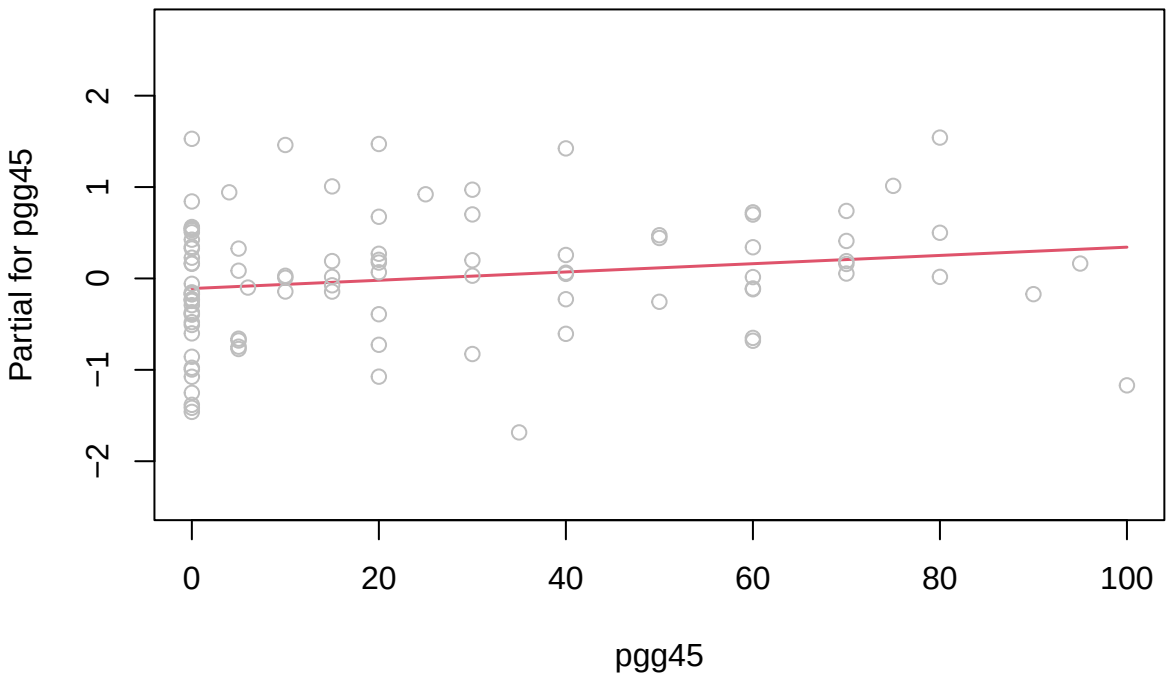

Partial for age
2
1
0
−1
−2
40
50
60
70
80
age


Partial for lbph
2
1
0
−1
−2
−1
0
1
2
lbph


Partial for svi
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
svi


Partial for lcp
2
1
0
−1
−2
−1
0
1
2
3
lcp


Partial for gleason
2
1
0
−1
−2
6.0
6.5
7.0
7.5
8.0
8.5
9.0
gleason



The flatness of the lines associated with all except for [lcavol, lweight] reflects their lack of significance. The model may be improved by dropping them.

6.5

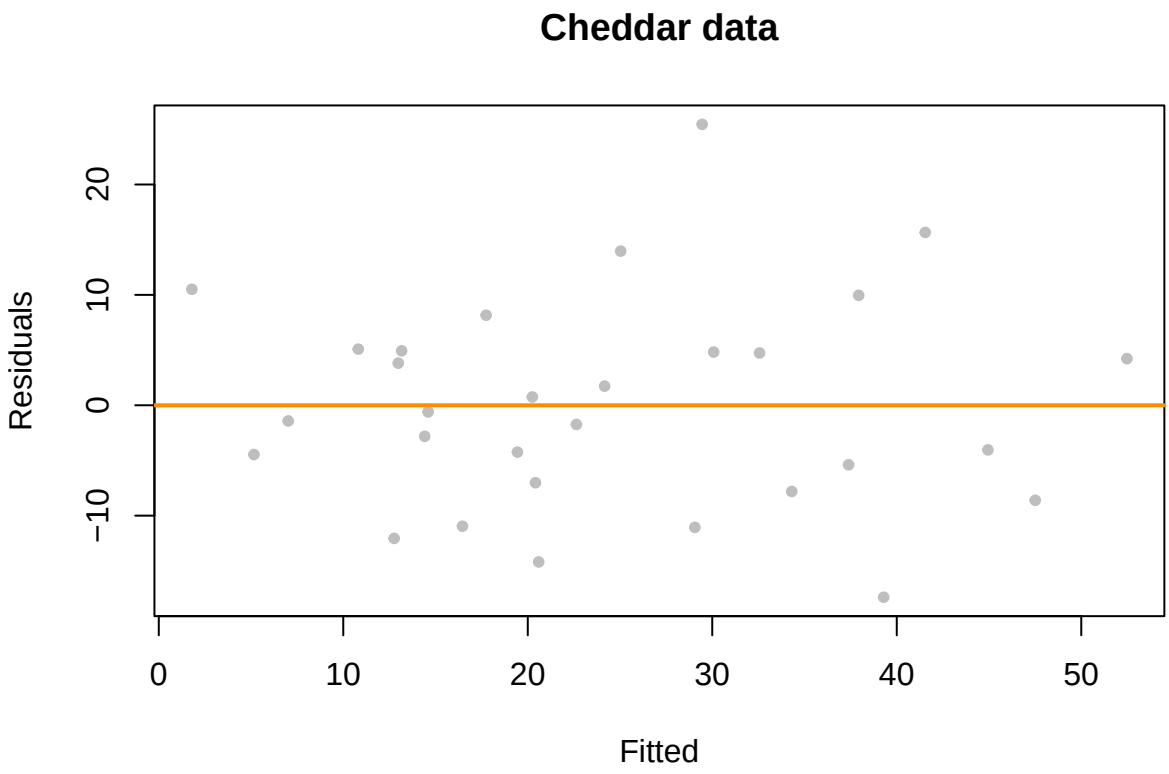
```
data("cheddar")
```

```
m <- lm(taste ~ ., data=cheddar)
```

a.

```
residuals <- resid(m)
```

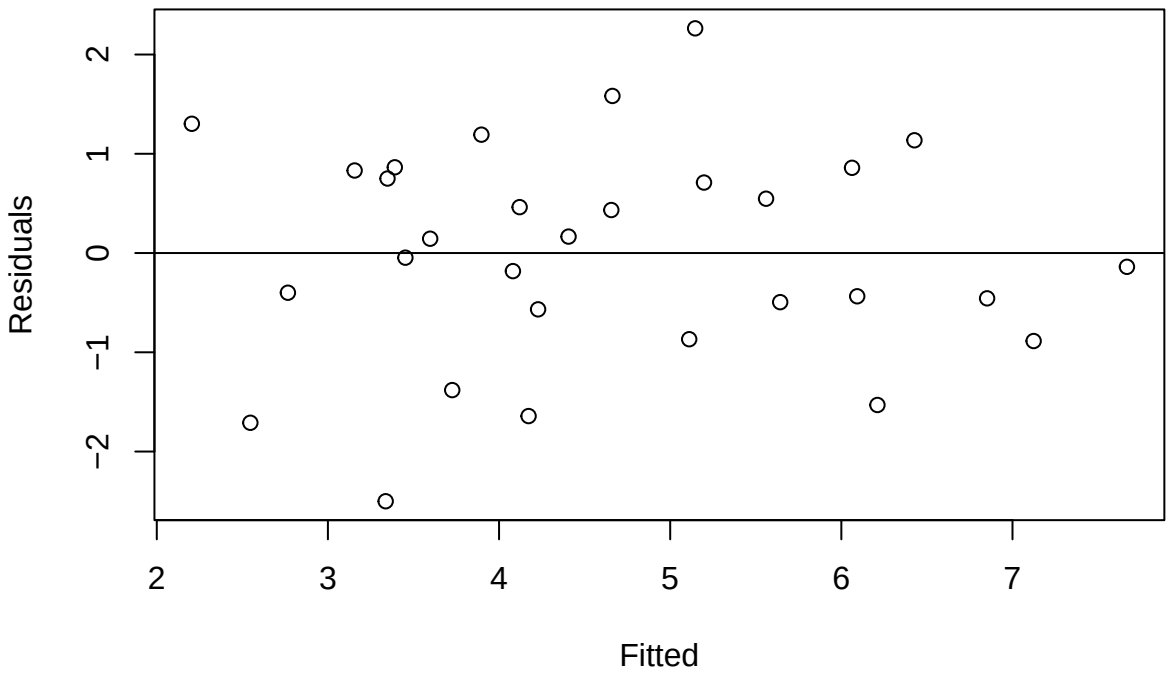
```
plot(fitted(m), residuals, col="grey", pch=20, xlab="Fitted", ylab="Residuals", main = "Cheddar data")
abline(h=0, col="darkorange", lwd=2)
```



From the residual-fitted plot, spread of residuals is constant across the range of fitted values, hence the equal variance assumption is satisfied. In addition, the plot indicates some non-linearity.

We can improve non-linearity by taking the sqrt.

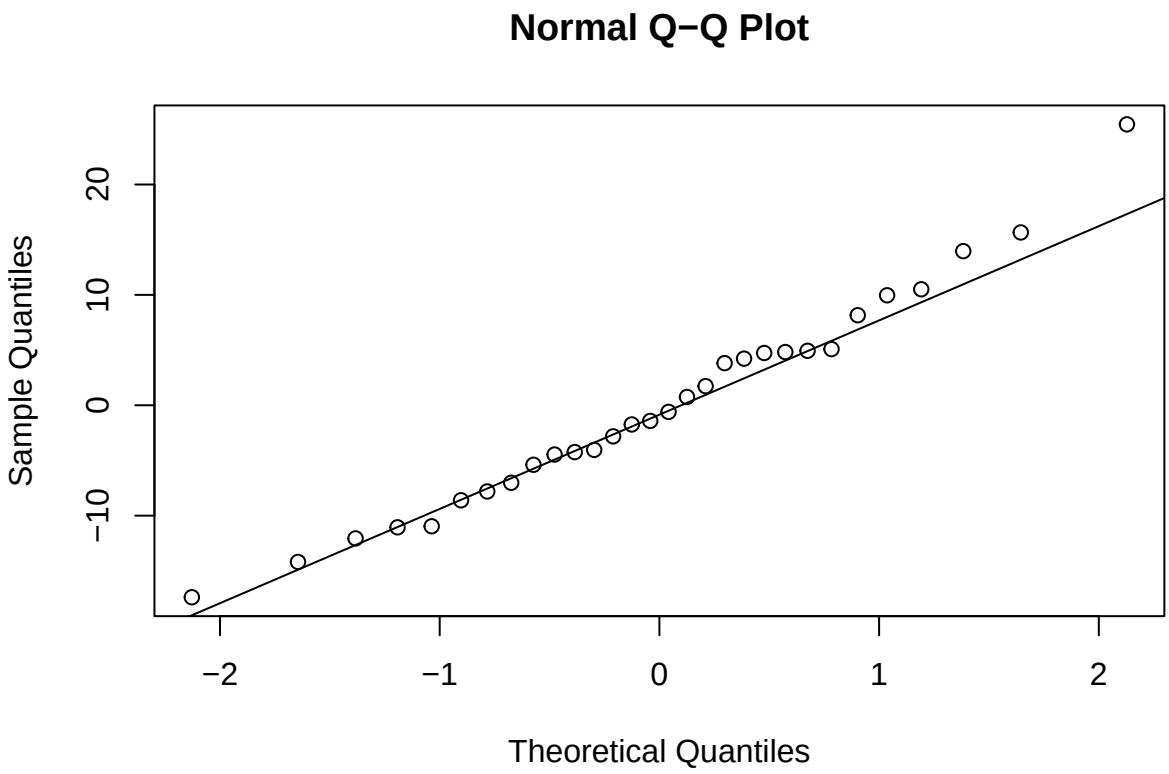
```
sqrt_m <- lm(sqrt(taste) ~ ., data=cheddar)
plot(fitted(sqrt_m), residuals(sqrt_m), xlab = "Fitted", ylab = "Residuals")
abline(h=0)
```



As variances hover more and more around 0, the non-linearity has been improved.

b.

```
qqnorm(residuals)
qqline(residuals)
```




```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.98021, p-value = 0.8312
```

Since the shapiro p-value is more than 0.05, we fail the reject the null hypothesis of the test which is the residuals are normally distributed. Hence the normally assumption is satisfied. Hence the normally assumption is satisfied, which is also supported by the QQ plot.

c.

```
hatvalues(m)[hatvalues(m) > 2 * mean(hatvalues(m))]
```

```
## named numeric(0)
```

We can see that there are 0 leverage points.

d.

```r
rstandard(m)[abs(rstandard(m)) > 2]
```

```
##       15
## 2.633351
```

We can see that observation [15] is an outlier.

e.

```r
cook_dist <- cooks.distance(m)
```

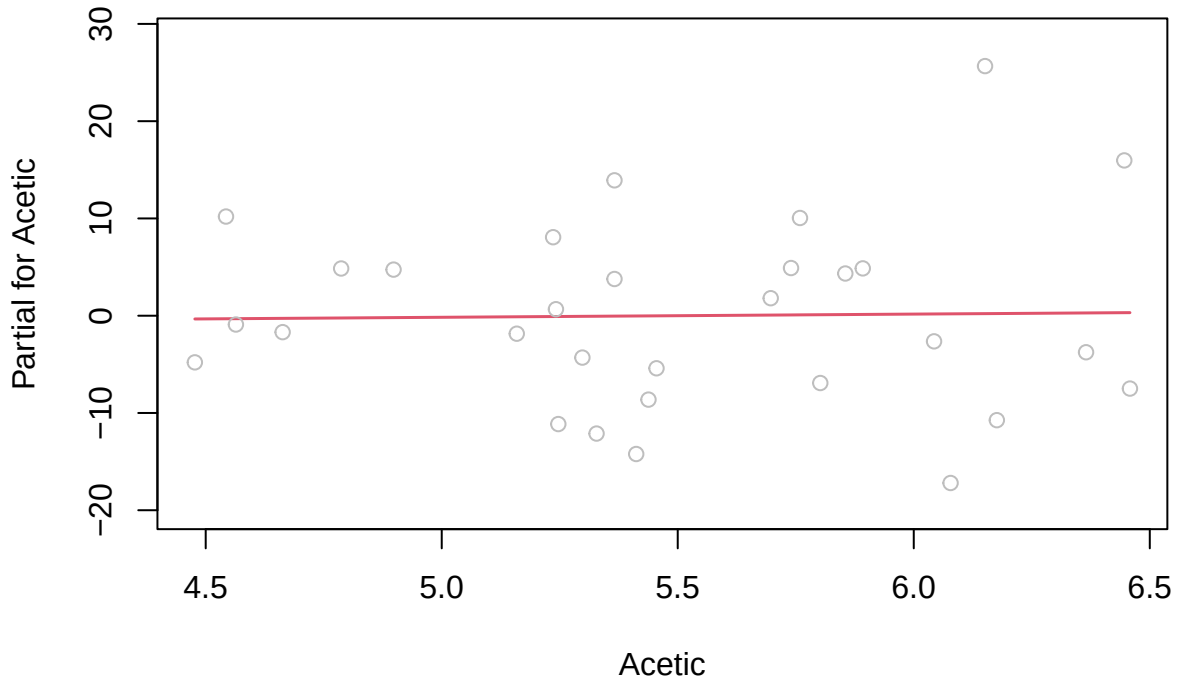
```r
cook_dist[15] > 4 / length(cook_dist)
```

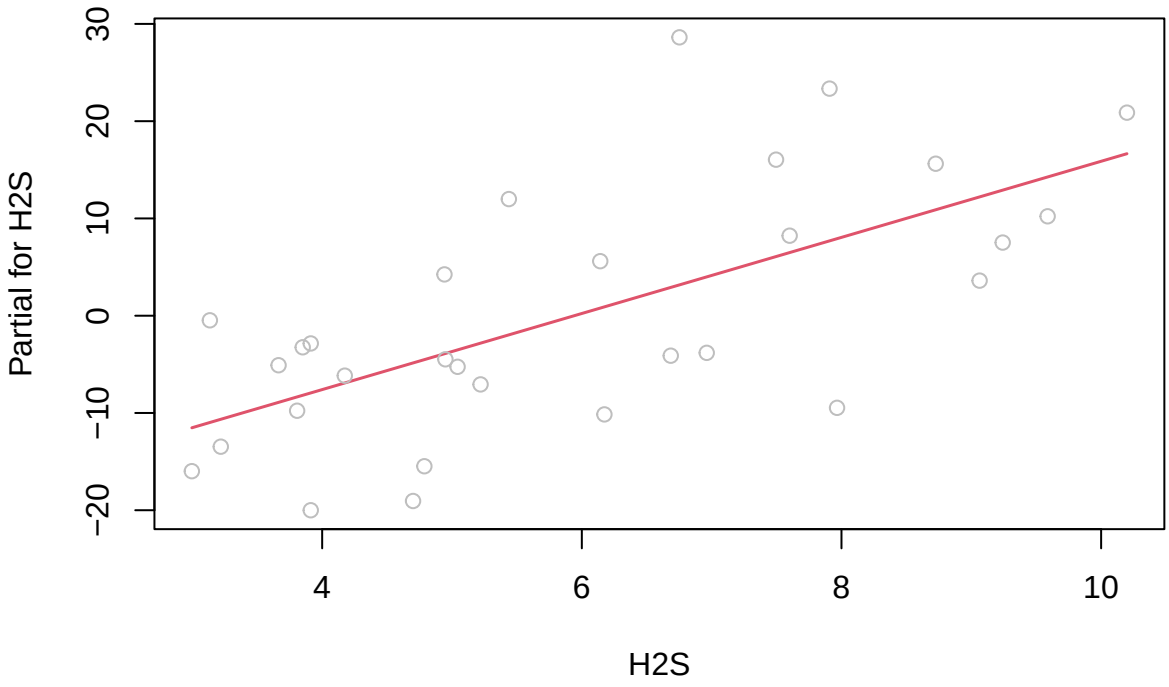
```
##   15
## TRUE
```

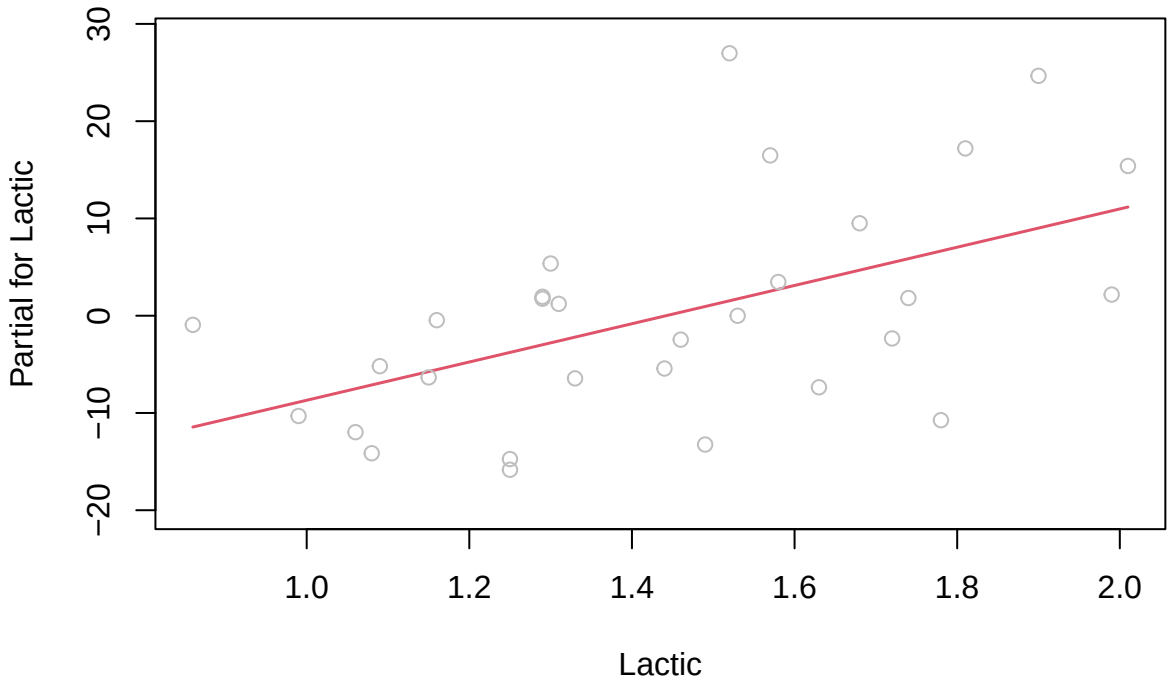
Observation [15] is an outlier and influential point, thus it should be dropped.

f.

```r
termplot(m,partial.resid = T)
```




Partial for H2S
30
20
10
0
−10
−20
4
6
8
10
H2S



The flatness of the lines associated with [Acetic] reflects their lack of significance.

The model may be improved by dropping it.

6.7

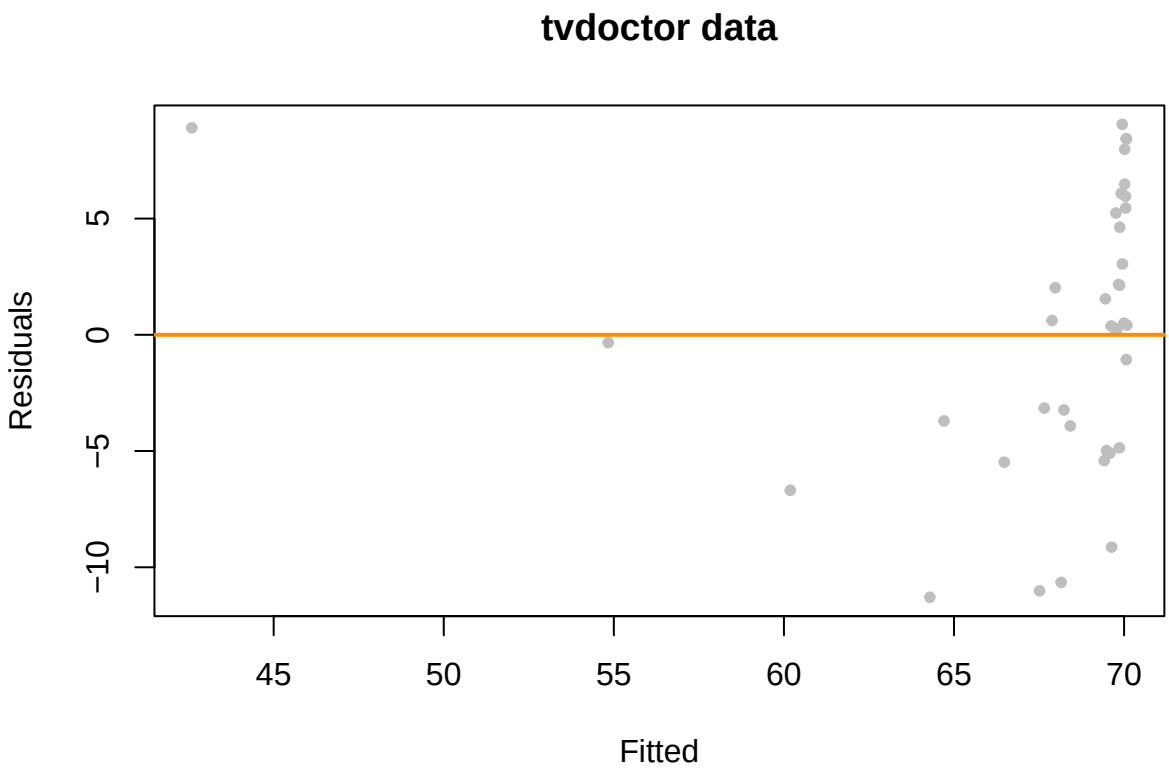
```
data("tvdoctor")
```

```
m <- lm(life ~ ., data=tvdoctor)
```

a.

```
residuals <- resid(m)
```

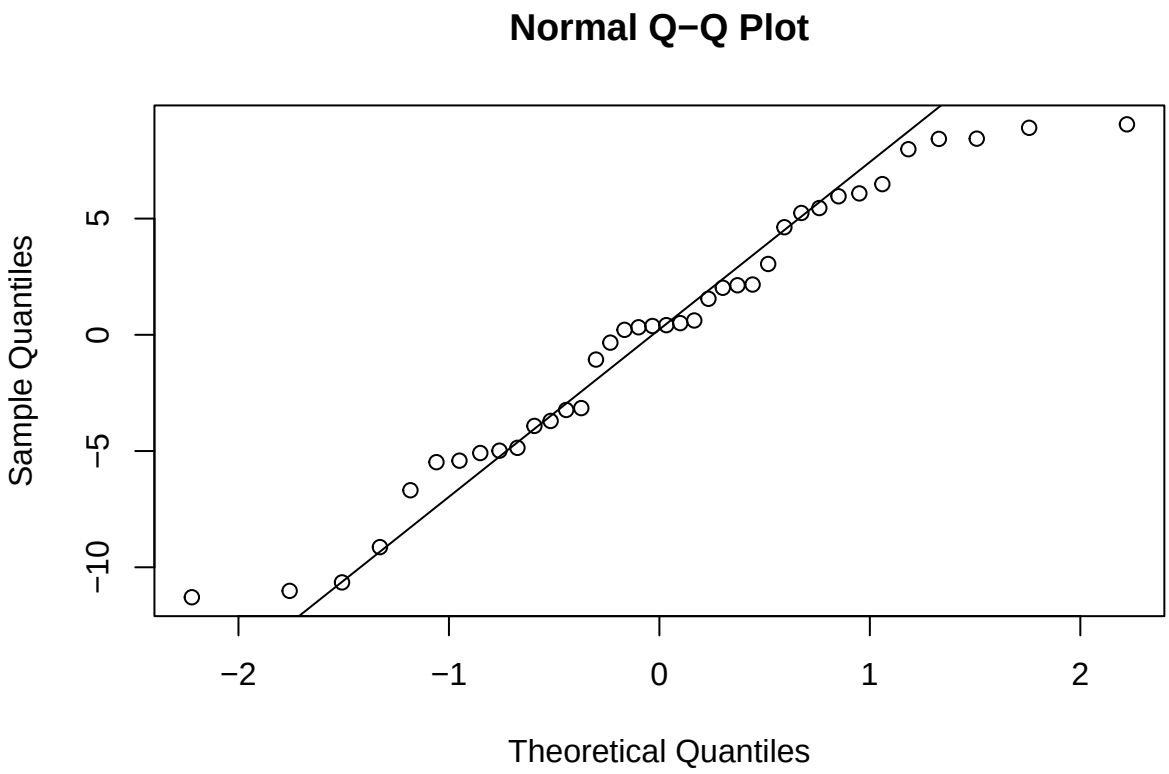
```
plot(fitted(m), residuals, col="grey", pch=20, xlab="Fitted", ylab="Residuals", main = "tvdoctor data")
abline(h=0, col="darkorange", lwd=2)
```



From the residual-fitted plot, spread of residuals is NOT constant across the range of fitted values, hence the equal variance assumption is violated In addition, the plot indicates lots of non-linearity and require a transformation of data.

b.

```
qqnorm(residuals)
qqline(residuals)
```




```
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.95872, p-value = 0.1725
```

Since the shapiro p-value is more than 0.05, we fail the reject the null hypothesis of the test which is the residuals are normally distributed. Hence the normally assumption is satisfied.

c.

```
hatvalues(m)[hatvalues(m) > 2 * mean(hatvalues(m))]
```

```
## Bangladesh   Ethiopia    Myanmar
##  0.1597777  0.8222873  0.7598006
```

We can see that [Bangladesh, Ethiopia, and Myanmar] are leverage points.

d.

```r
rstandard(m)[abs(rstandard(m)) > 2]
```

```
##  Ethiopia     Sudan
##  3.518939 -2.042465
```

We can see that [Ethiopia, Sudan] are outliers.

Since Ethiopia appeared in both leverage points and outliers, it has a significant pull to our linear model. Hence, it's better to remove it.

e.

```r
cook_dist <- cooks.distance(m)
```

```r
cook_dist['Ethiopia'] > 4 / length(cook_dist)
```

```
## Ethiopia
##     TRUE
```

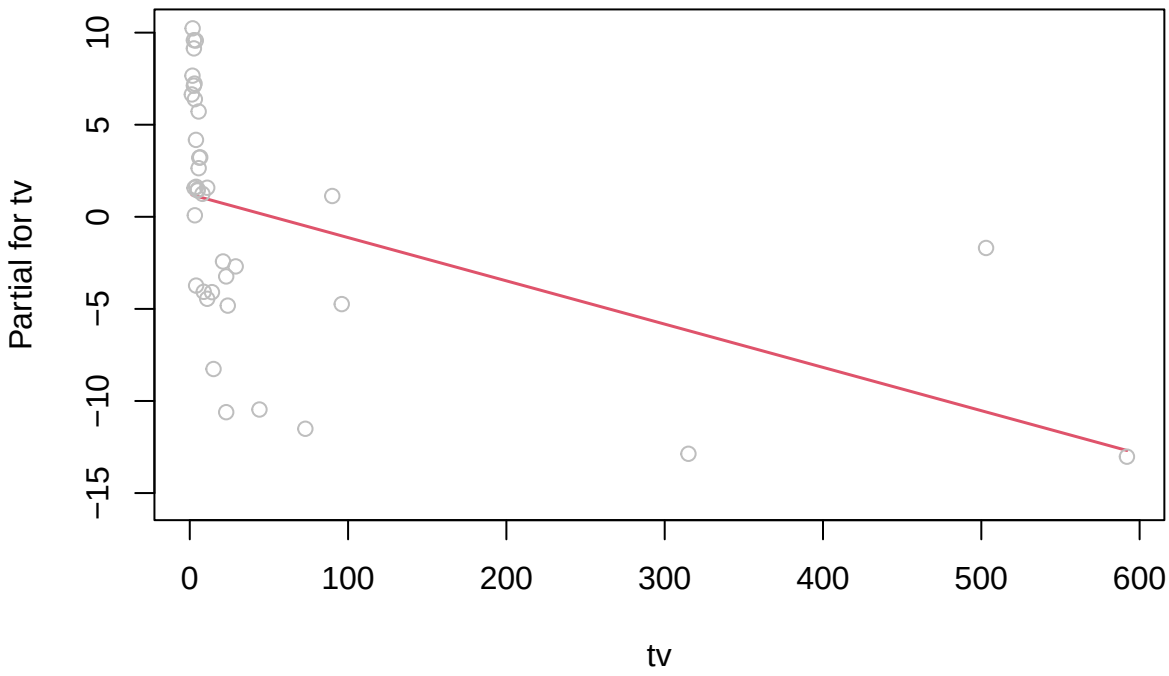
```r
cook_dist['Sudan'] > 4 / length(cook_dist)
```

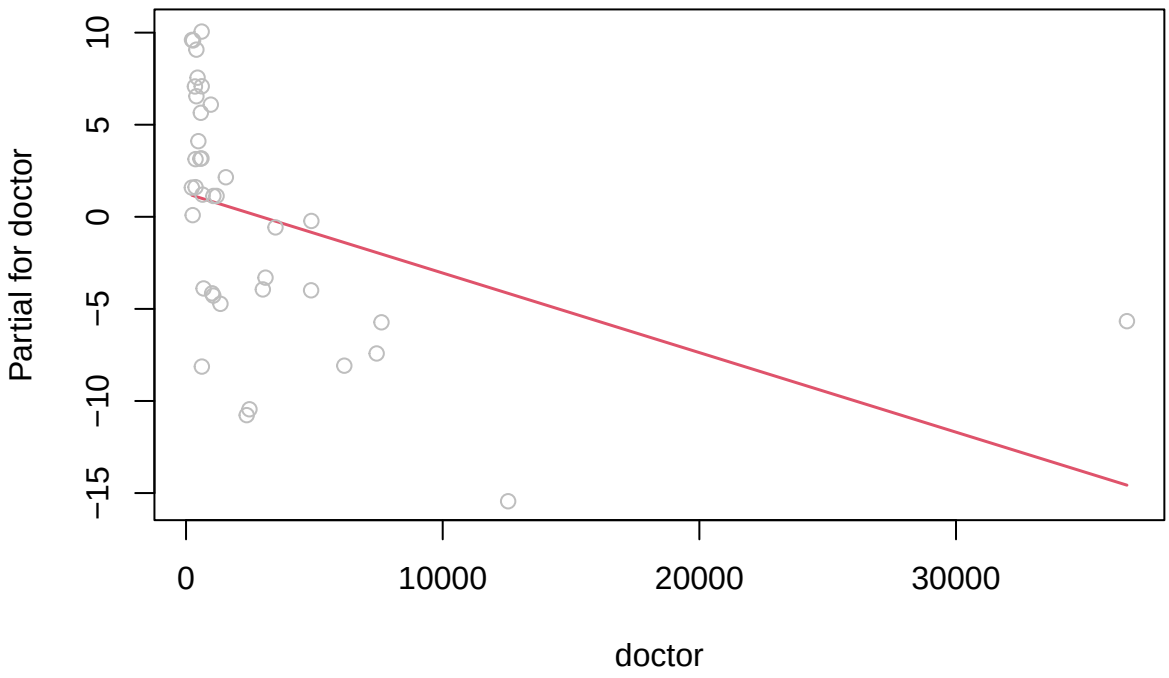
```
## Sudan
##  TRUE
```

We can see that both [Ethiopia, Sudan] are influential points, and should be removed.

f.

```r
termplot(m,partial.resid = T)
```


Partial for tv
10
5
0
−5
−10
−15
0
100
200
300
400
500
600
tv



There is no flatness of the lines hence all predictors are significant.