HW3

Tianqi Wu

3/9/2020

```
library(faraway)
library(lmtest)
library(car)
attach(sat)
```

Problem 1

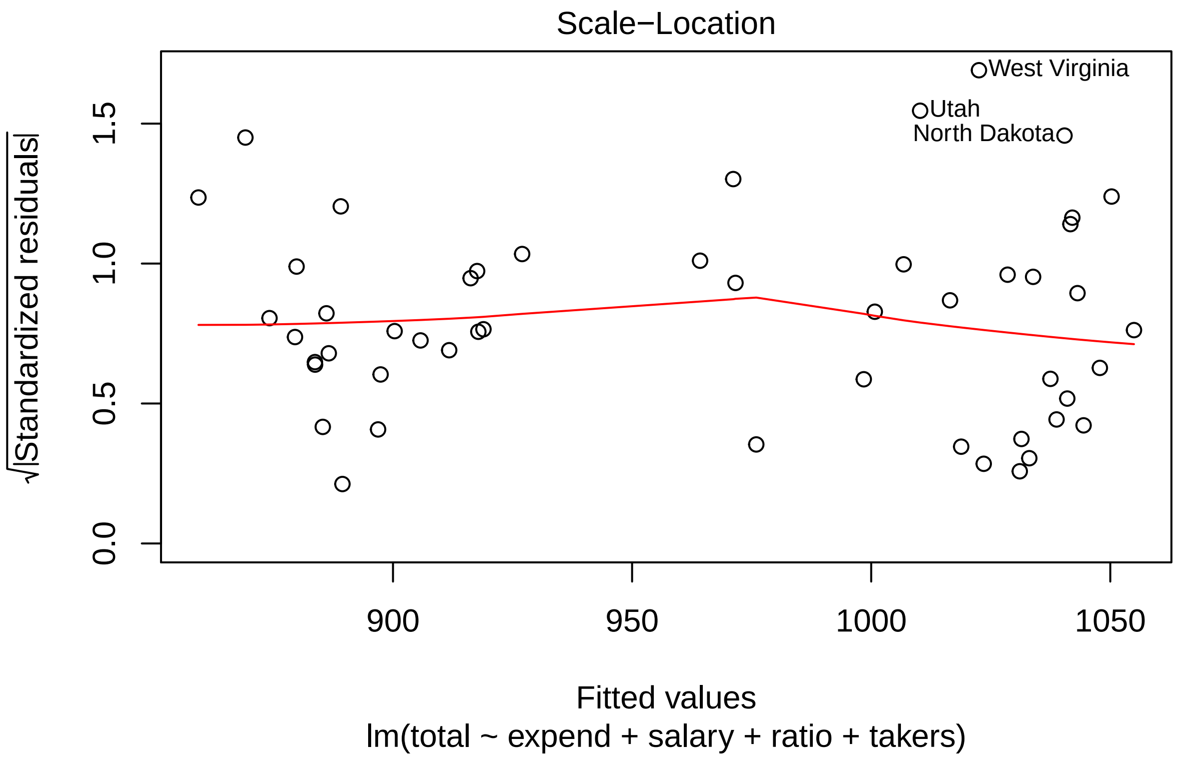
(a)

From the scale-location plot, we can see a horizontal line with equally spread points, which indicates constant variance. Also, from the Breusch-Pagan test, since the p-value is $0.7066 > 0.05$. We fail to reject the null hypothesis and conclude that constant variance assumption for the errors is valid. West virginia, North Dakota and Utah deviate from the line and we may examine them more.

```
model.1 = lm(total~expend+salary+ratio+takers,data=sat)
bptest(model.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.1
## BP = 2.1587, df = 4, p-value = 0.7066
```

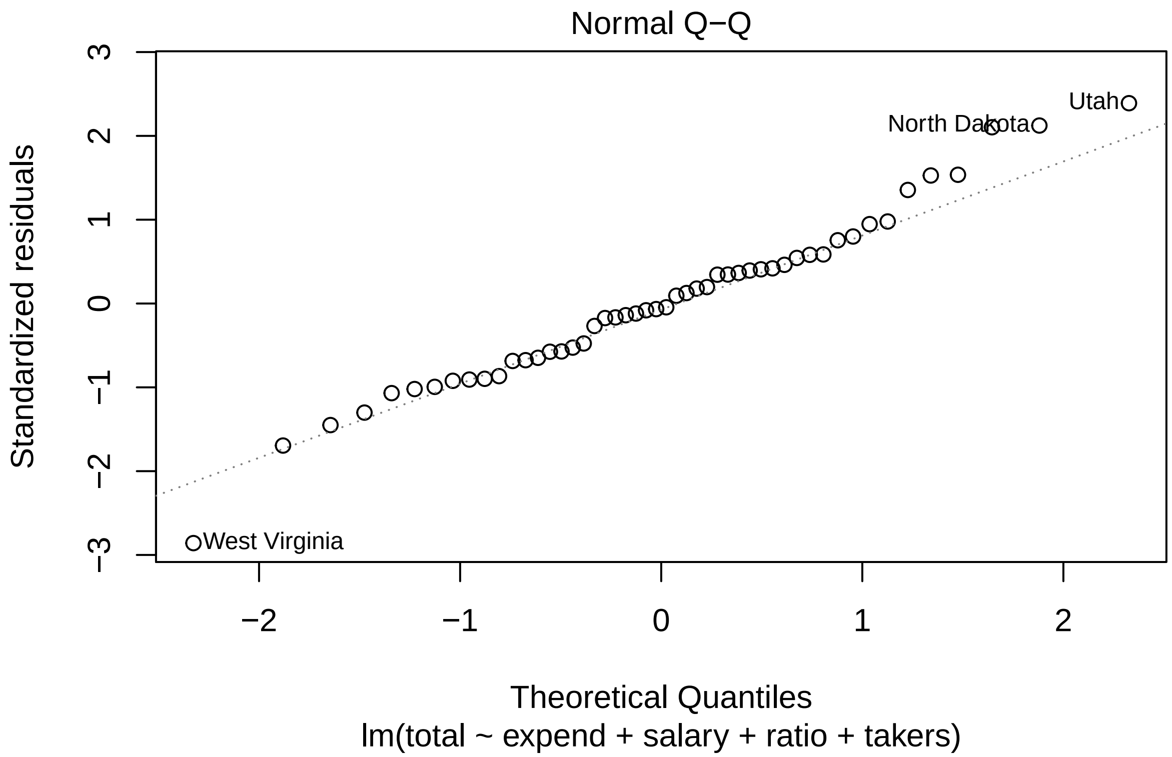
```
plot(model.1,which=3)
```



(b)

From the qq-plot, since it is approximately in a line, the normality assumption is valid. West virginia, North Dakota and Utah deviate from the line and we may examine them more. Also, from Shapiro-Wilk normality test, since p-value is 0.4304 > 0.05, We fail to reject the null hypothesis and conclude that normality assumption is valid.

```
plot(model.1,which=2)
```



```
shapiro.test(residuals(model.1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.1)
## W = 0.97691, p-value = 0.4304
```

(c)

From the anlaysis, California, Connecticut, New Jersey and Utah are large leverage points. We may examine them closely.

```
lev=influence(model.1)$hat
n=nrow(sat);p=5
sat[lev > 2*p/n,]
```

```
##               expend ratio salary takers verbal math total
## California   4.992  24.0 41.078     45    417  485   902
## Connecticut  8.817  14.4 50.045     81    431  477   908
## New Jersey   9.774  13.8 46.087     70    420  478   898
## Utah         3.656  24.3 29.082      4    513  563  1076
```

(d)

West virginia has largest student residual 3.124 $<$ 3.525. Hence, at 0.05 significance level, there is no outlier.

```
jack=rstudent(model.1);
qt(.05/(2*n), n-p-1) # Bonferroni correction
```

```
## [1] -3.525801
```

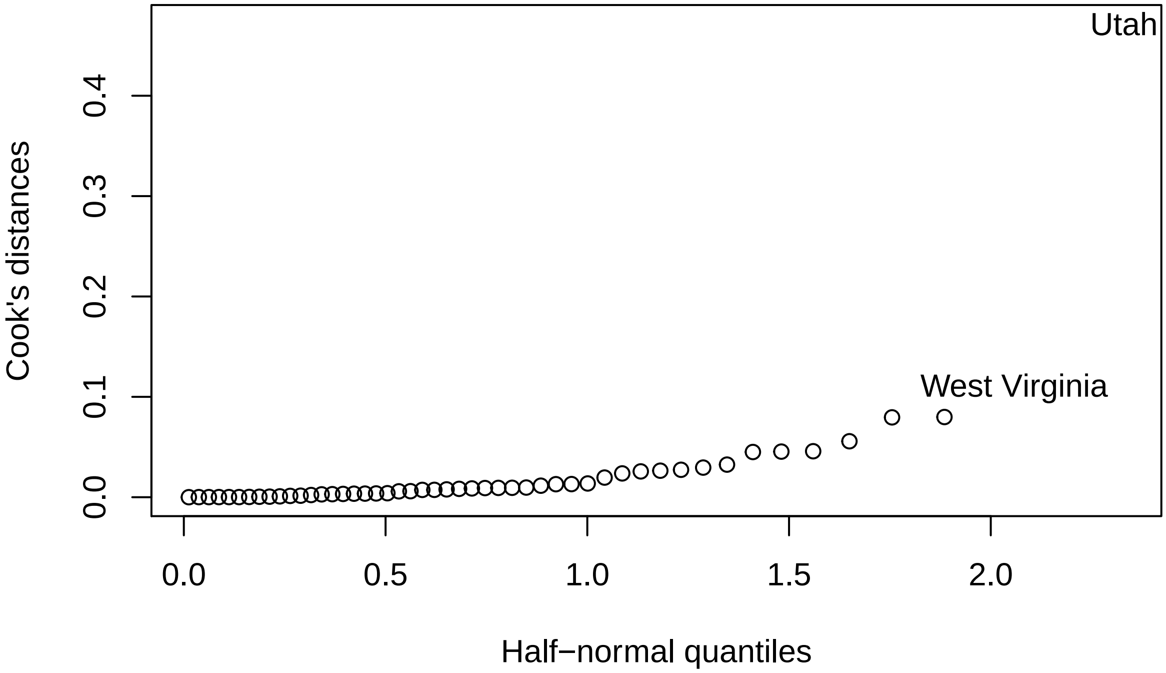
```
sort(abs(jack), decreasing=TRUE)[1:5]
```

```
## West Virginia          Utah  North Dakota New Hampshire        Nevada
##      3.124428      2.529587      2.213686      2.190006      1.732004
```

(e)

Utah has max cook's distance 0.4715 < 1. Although there are no high influential points based on the rule-of-thumb, the cook's distance for Utah is much larger than the other samples. So, we may remove "Utah", refit the model, and check the changes.

```
cook = cooks.distance(model.1)
halfnorm(cook, labs=row.names(sat), ylab="Cook's distances")
```



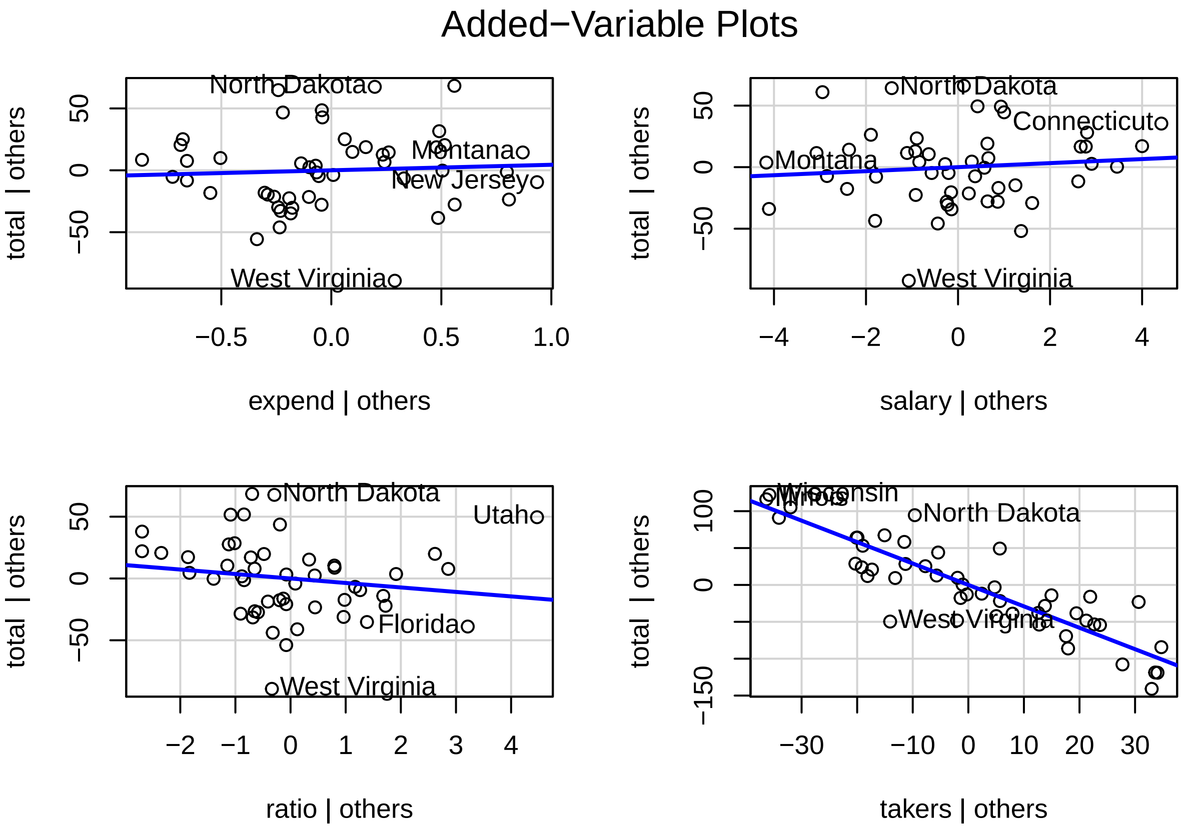
```
max(cook)
```

```
## [1] 0.4715287
```

(f)

Since the added-variable plot produce points randomly scattered around a line through the origin for all the variables, the linear model structure assumption is valid. We also observe that "total" residuals decrease as takers residuals increase.

```
avPlots(model.1)
```



Problem2

(a)

From the scale-location plot, we can see the line has some nonlinearity. Also, the variance gets larger as fitted value increases. There is evidence of heterokedasticity and the constant variance assumption may not be valid. However, from the Breusch-Pagan test, since the p-value is 0.1693 $>$ 0.05. We fail to reject the null hypothesis of homocedasticity. Oberservation 24,36 and 39 deviate from the line and we may examine them more.

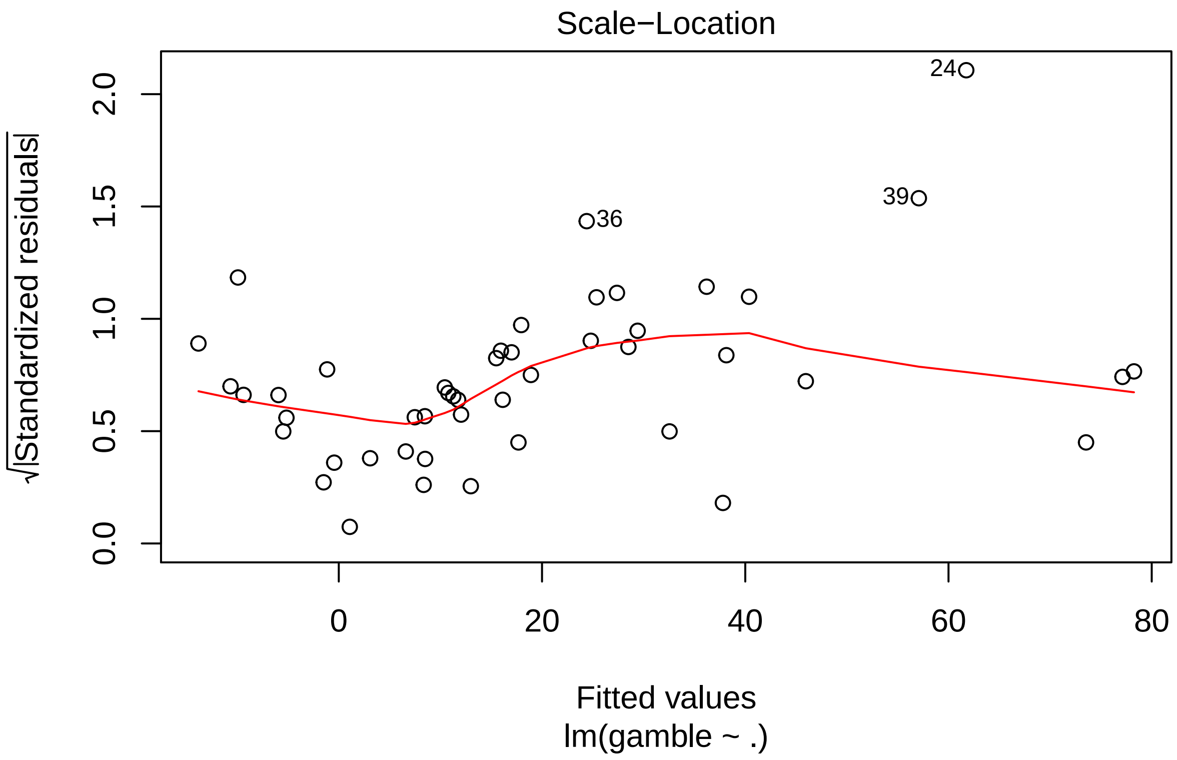
```
attach(teengamb)
```

```
## The following object is masked from sat:
##
##     verbal
```

```
model.2 = lm(gamble~.,data=teengamb)
bptest(model.2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.2
## BP = 6.4288, df = 4, p-value = 0.1693
```

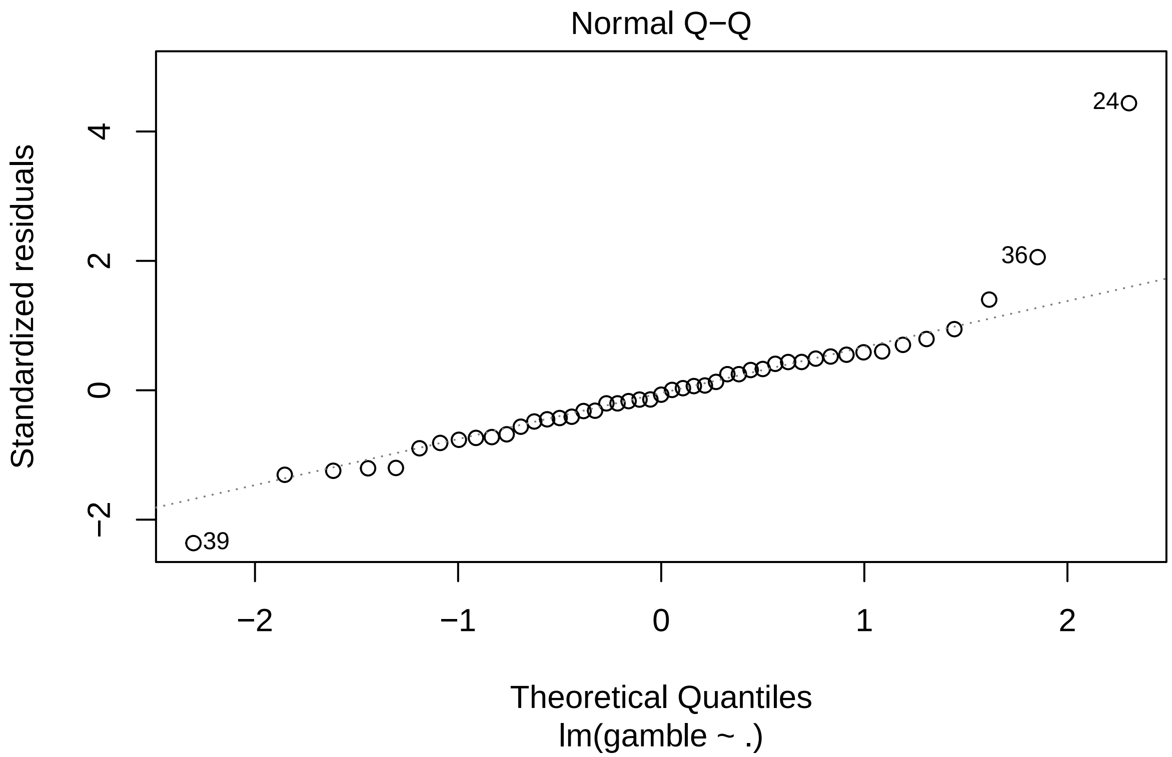
```
plot(model.2,which=3)
```



(b)

From the qq-plot, we find that data may have heavy-tail problem and the normality assumption may not be valid. Oberservation 24,36 and 39 deviate from the line and we may examine them more. From the Shapiro-Wilk normality test, since p-value is 8.16e-05 < 0.05, we reject the null hypothesis and conclude that normality assumption is not valid. To solve the problem, we may use tranformation, robust regression or permutation test which does not have normality assumption.

```
plot(model.2,which=2)
```



```
shapiro.test(residuals(model.2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.2)
## W = 0.86839, p-value = 8.16e-05
```

(c)

From the anlaysis, 31,33,35 and 42 are large leverage points and we may examine them more.

```
lev.2=influence(model.2)$hat
n.2=nrow(teengamb);p.2=ncol(teengamb)
teengamb[lev.2 > 2*p.2/n.2,]
```

```
##    sex status income verbal gamble
## 31   0     18   12.0      2   88.0
## 33   0     38   15.0      7   90.0
## 35   0     28    1.5      1   14.1
## 42   0     61   15.0      9   69.7
```

(d)

Oberservation 24 has largest student residual 6.016116 > 3.522. Hence, at 0.05 significance level, oberservation 24 is the outlier.

```
jack.2=rstudent(model.2);
qt(.05/(2*n.2), n.2-p.2-1) # Bonferroni correction
```

```
## [1] -3.522795
```

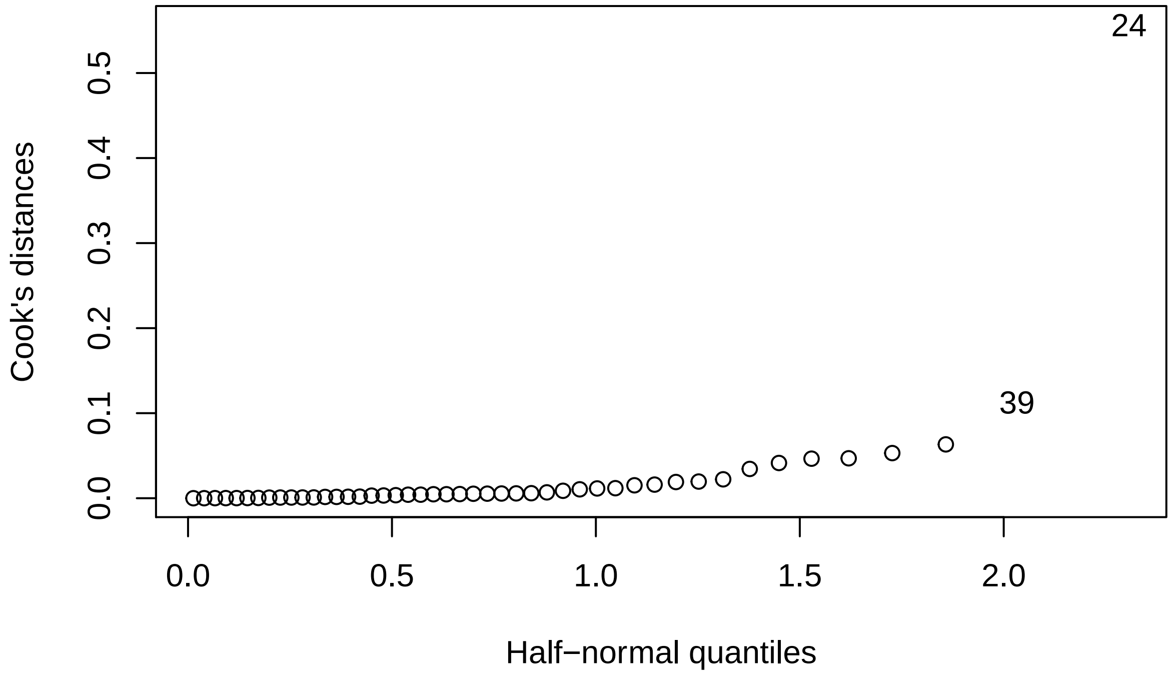
```
sort(abs(jack.2), decreasing=TRUE)[1:5]
```

```
##       24       39       36        5       18
## 6.016116 2.506090 2.144826 1.418583 1.317398
```

(e)

Oberservation 24 has max cook's distance 0.5565 < 1. Although there are no high influential points based on the rule-of-thumb, the cook's distance for oberservation 24 is much larger than the other samples. So, we may remove it, refit the model, and check the changes.

```
cook.2 = cooks.distance(model.2)
halfnorm(cook.2, labs=row.names(teengamb), ylab="Cook's distances")
```



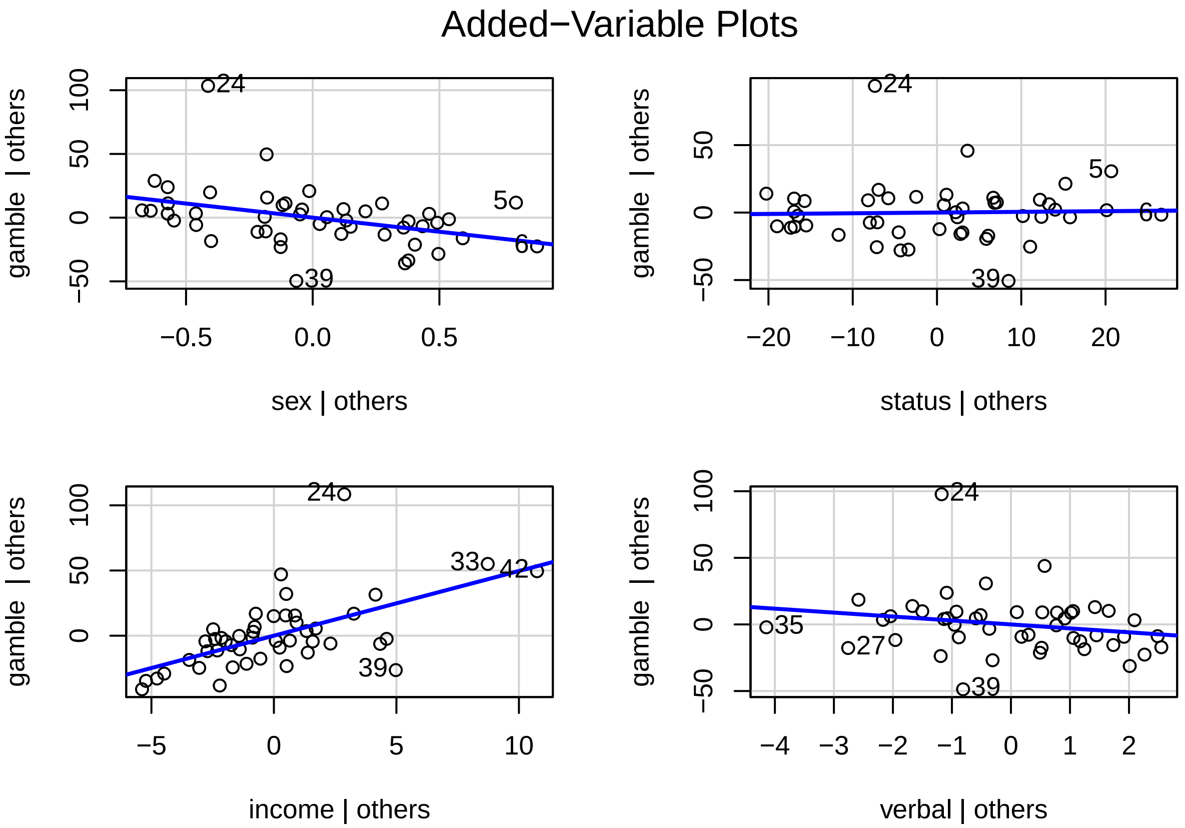
```
max(cook.2)
```

```
## [1] 0.5565011
```

(f)

Since the added-variable plot produce points randomly scattered around a line through the origin for all the variables, the linear model structure assumption is valid. In every plot, observation 24 deviate from the line and we may remove it. We also oberserve that gamble residuals increase as income residuals increase.

```
avPlots(model.2)
```



Problem 3

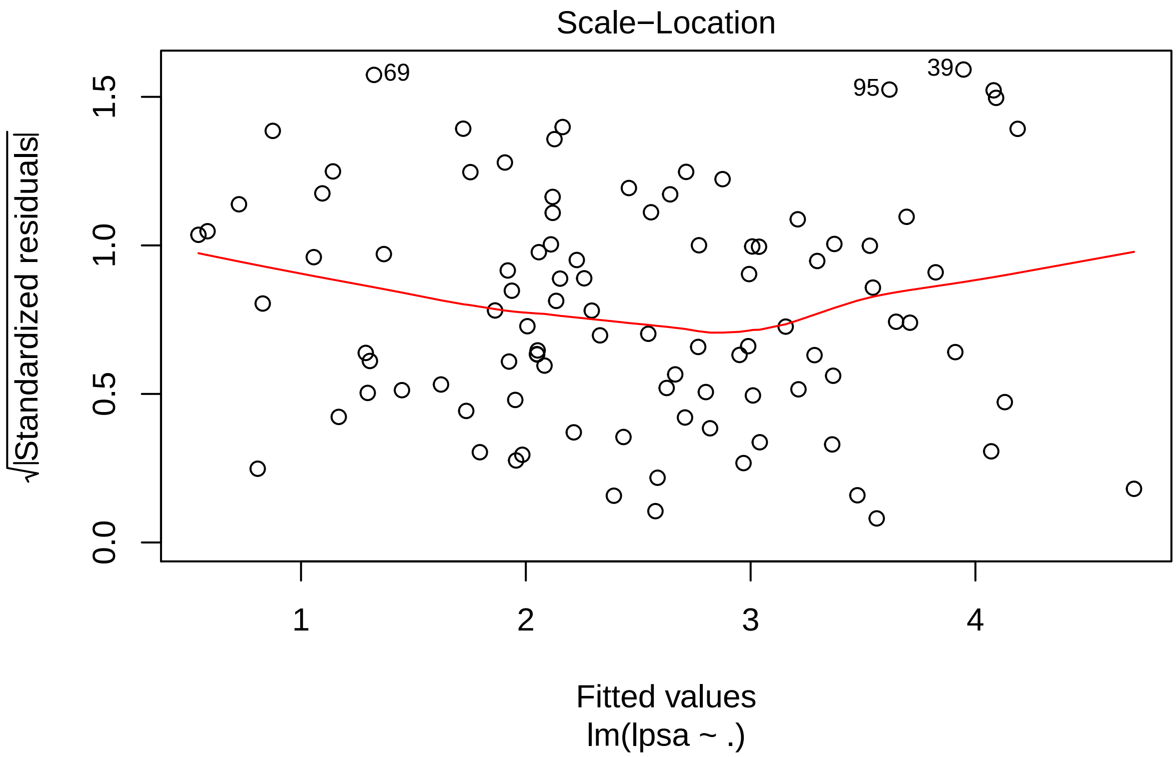
(a)

From the scale-location plot, we can see a nearly horizontal line with equally spread points and there is no obvious pattern. The scale-location plot indicates constant variance. Also, from the Breusch-Pagan test, since the p-value is 0.2594 > 0.05. We fail to reject the null hypothesis and conclude that constant variance assumption for the errors is valid. Oberservation 39,69 and 95 deviate from the line and we may examine them more.

```r
attach(prostate)
model.3 = lm(lpsa~.,data=prostate)
bptest(model.3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.3
## BP = 10.08, df = 8, p-value = 0.2594
```

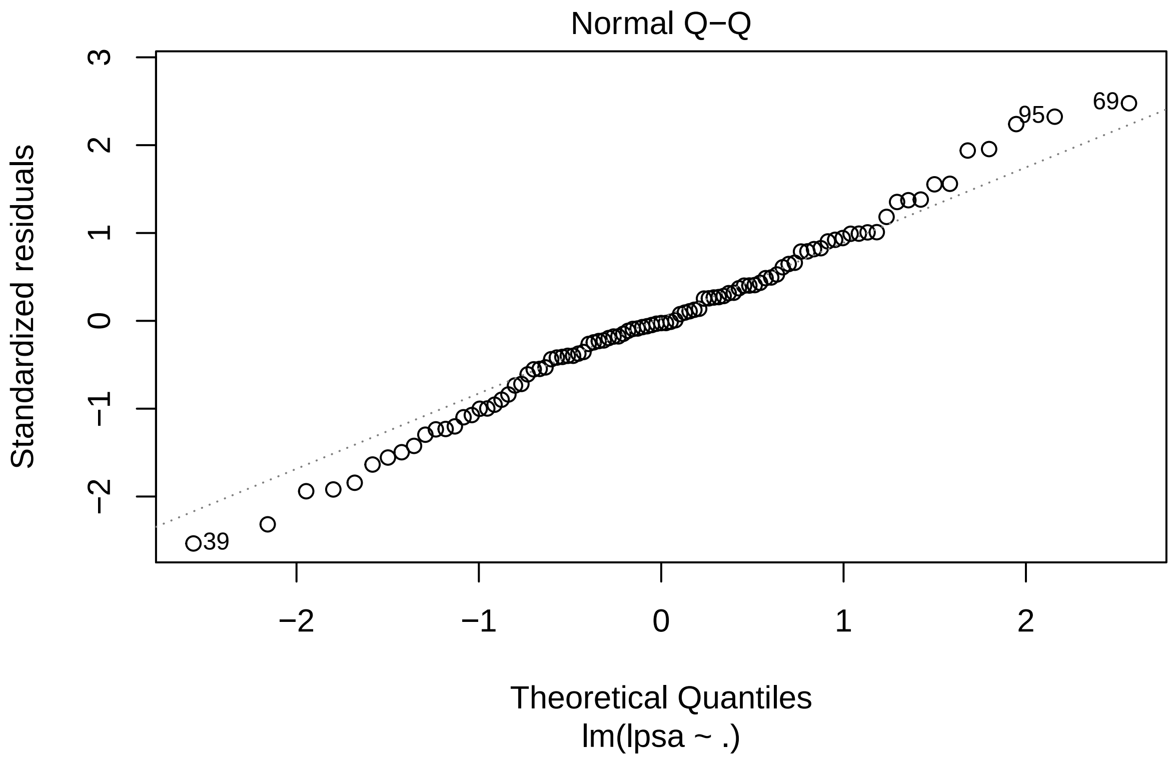
```
plot(model.3,which=3)
```



(b)

From the qq-plot, since it is approximately in a line, the normality assumption is valid. Oberservation 39,69 and 95 deviate from the line and we may examine them more. Also, from Shapiro-Wilk normality test, since p-value is 0.7721 > 0.05, We fail to reject the null hypothesis and conclude that normality assumption is valid.

```
plot(model.3,which=2)
```



```
shapiro.test(residuals(model.3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.3)
## W = 0.99113, p-value = 0.7721
```

(c)

From the anlaysis, 32,37,41,74,92 are large leverage points and we may examine them more.

```
lev.3=influence(model.3)$hat
n.3=nrow(prostate);p.3=ncol(prostate)
prostate[lev.3 > 2*p.3/n.3,]
```

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45    lpsa
## 32 0.1823216  6.1076  65  1.704748   0 -1.38629       6     0 2.00821
## 37 1.4231083  3.6571  73 -0.579818   0  1.65823       8    15 2.15756
## 41 0.6205765  3.1420  60 -1.386294   0 -1.38629       9    80 2.29757
## 74 1.8389611  3.2367  60  0.438255   1  1.17865       9    90 3.07501
## 92 2.5329028  3.6776  61  1.348073   1 -1.38629       7    15 4.12955
```

(d)

Oberservation 39 has largest student residual 2.616980 $<$ 3.607426. Hence, at 0.05 significance level, there is no outlier.

```
jack.3=rstudent(model.3);
qt(.05/(2*n.3), n.3-p.3-1) # Bonferroni correction
```

```
## [1] -3.607426
```

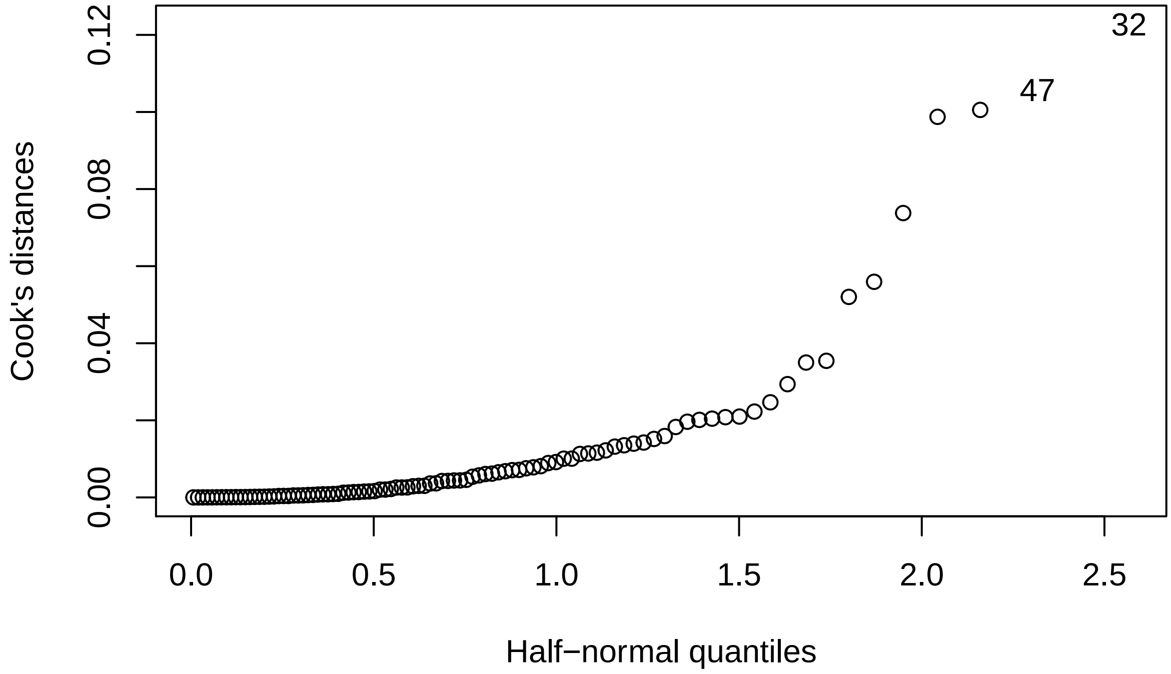
```
sort(abs(jack.3), decreasing=TRUE)[1:5]
```

```
##       39       69       95       47       97
## 2.616980 2.553530 2.385070 2.376671 2.293279
```

(e)

Oberservation 32 has max cook's distance 0.1226977 < 1. Although there are no high influential points based on the rule-of-thumb, the cook's distance for oberservation 32 is much larger than the other samples. So, we may remove it, refit the model, and check the changes.

```
cook.3 = cooks.distance(model.3)
halfnorm(cook.3, labs=row.names(teengamb), ylab="Cook's distances")
```



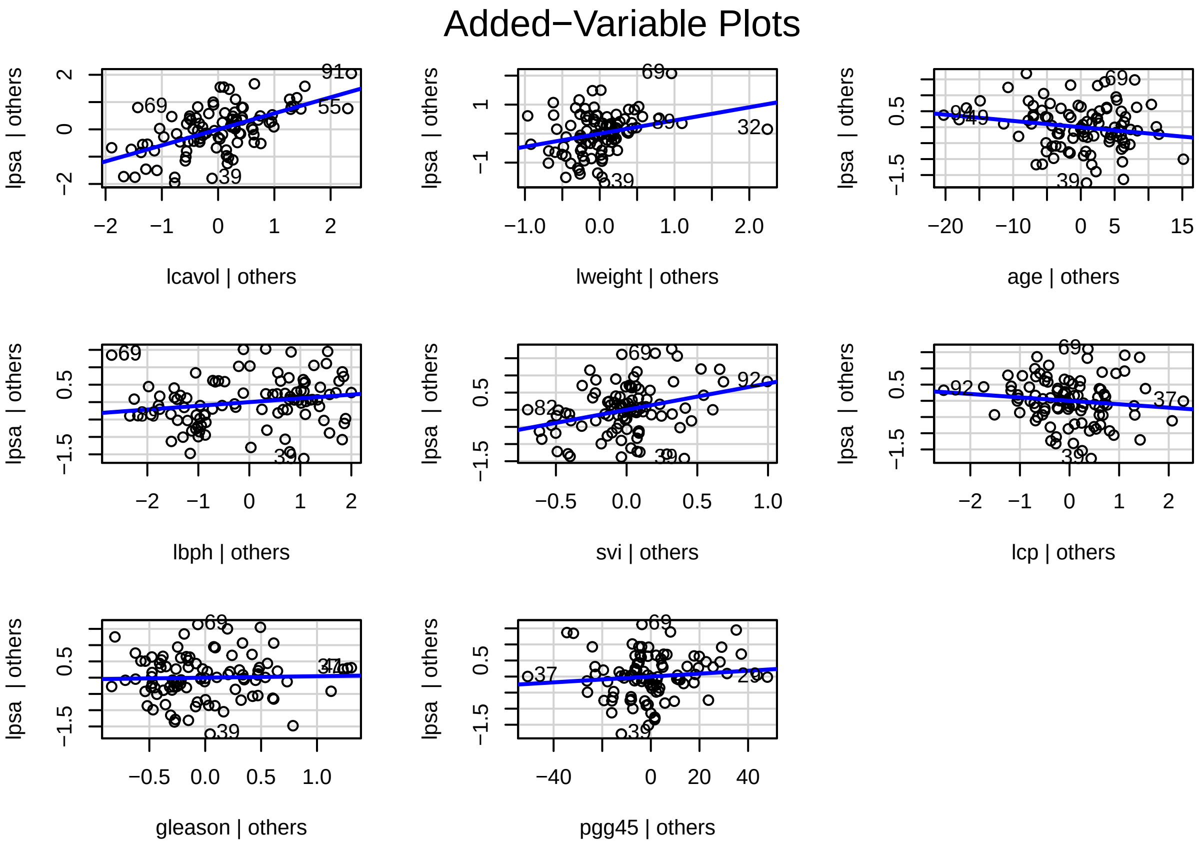
```
max(cook.3)
```

```
## [1] 0.1226977
```

(f)

Since the added-variable plot produce points randomly scattered around a line through the origin for all the variables, the linear model structure assumption is valid. We also oberserve that lpsa residuals increase as lcavol residuals increase.

```
avPlots(model.3)
```


Added−Variable Plots
lpsa | others
lcavol | others
−2
−1
0
1
2
91
55
69
39
lweight | others
−1.0
0.0
1.0
2.0
69
89
32
39
age | others
−20
−10
0
5
15
0.5
−1.5
69
49
39
lbph | others
−2
−1
0
1
2
69
39
svi | others
−0.5
0.0
0.5
1.0
69
92
82
39
lcp | others
92
37
69
39
gleason | others
−0.5
0.0
0.5
1.0
69
37
41
39
pgg45 | others
−40
0
20
40
69
37
29
39