

HW3

Chapter 5, Exercise 1

Following chunk contains the code for this exercise 1. First the teengamb dataset is loaded in the R environment. Then the response is modeled with just sex as a predictor. Then other variables are successively added. Then the Coefficients and p-values of sex for each model are plotted to examine their variation. It is found that the coefficients varies between -22 and -36 and p-value between 0.0004 to 0.01. The coefficients remains negative throughout and the p-value remains below 0.05 throughout which indicates that the effect of sex on response is stable.

```
data(teengamb, package='faraway')
beta<-rep(1,8)
pval<-rep(1,8)

lmod1<-lm(gamble~sex, teengamb)
beta[1]<-summary(lmod1)$coef[2,1]
pval[1]<-summary(lmod1)$coef[2,4]

lmod2<-lm(gamble~sex+status, teengamb)
beta[2]<-summary(lmod2)$coef[2,1]
pval[2]<-summary(lmod2)$coef[2,4]

lmod3<-lm(gamble~sex+income, teengamb)
beta[3]<-summary(lmod3)$coef[2,1]
pval[3]<-summary(lmod3)$coef[2,4]

lmod4<-lm(gamble~sex+verbal, teengamb)
beta[4]<-summary(lmod4)$coef[2,1]
pval[4]<-summary(lmod4)$coef[2,4]

lmod5<-lm(gamble~sex+status+income, teengamb)
beta[5]<-summary(lmod5)$coef[2,1]
pval[5]<-summary(lmod5)$coef[2,4]

lmod6<-lm(gamble~sex+status+verbal, teengamb)
beta[6]<-summary(lmod6)$coef[2,1]
pval[6]<-summary(lmod6)$coef[2,4]

lmod7<-lm(gamble~sex+status+income, teengamb)
beta[7]<-summary(lmod7)$coef[2,1]
pval[7]<-summary(lmod7)$coef[2,4]

lmod8<-lm(gamble~sex+status+income+verbal, teengamb)
beta[8]<-summary(lmod8)$coef[2,1]
pval[8]<-summary(lmod8)$coef[2,4]

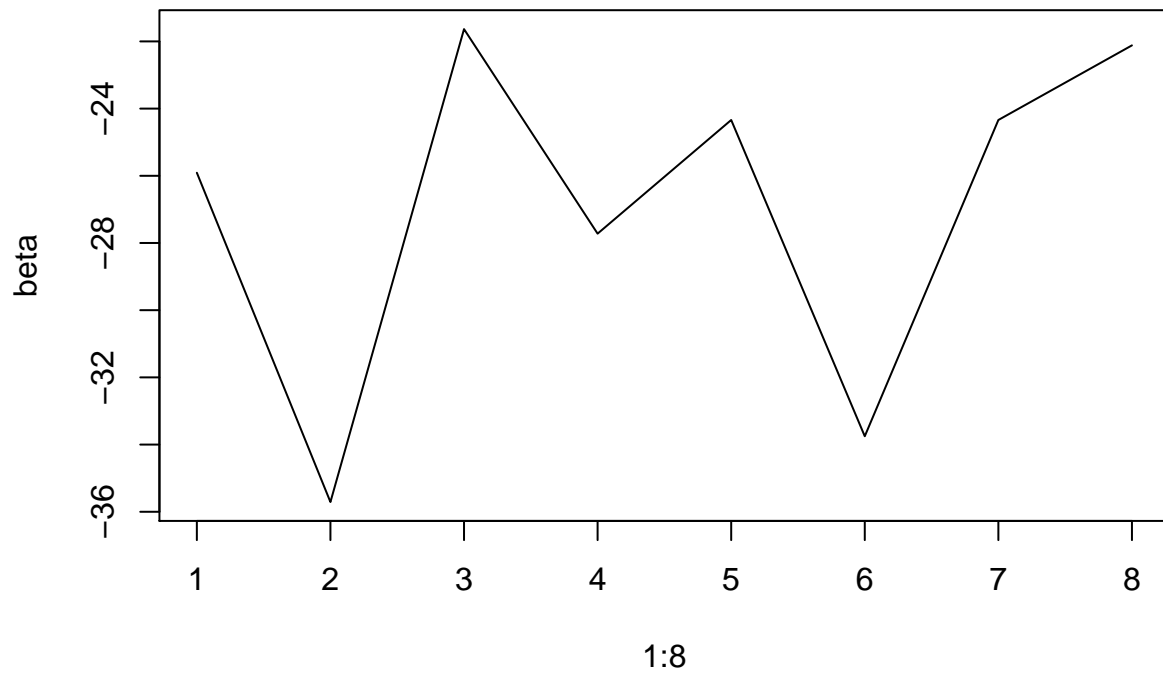
beta

## [1] -25.90921 -35.70937 -21.63439 -27.72208 -24.33934 -33.75202 -24.33934
## [8] -22.11833
```

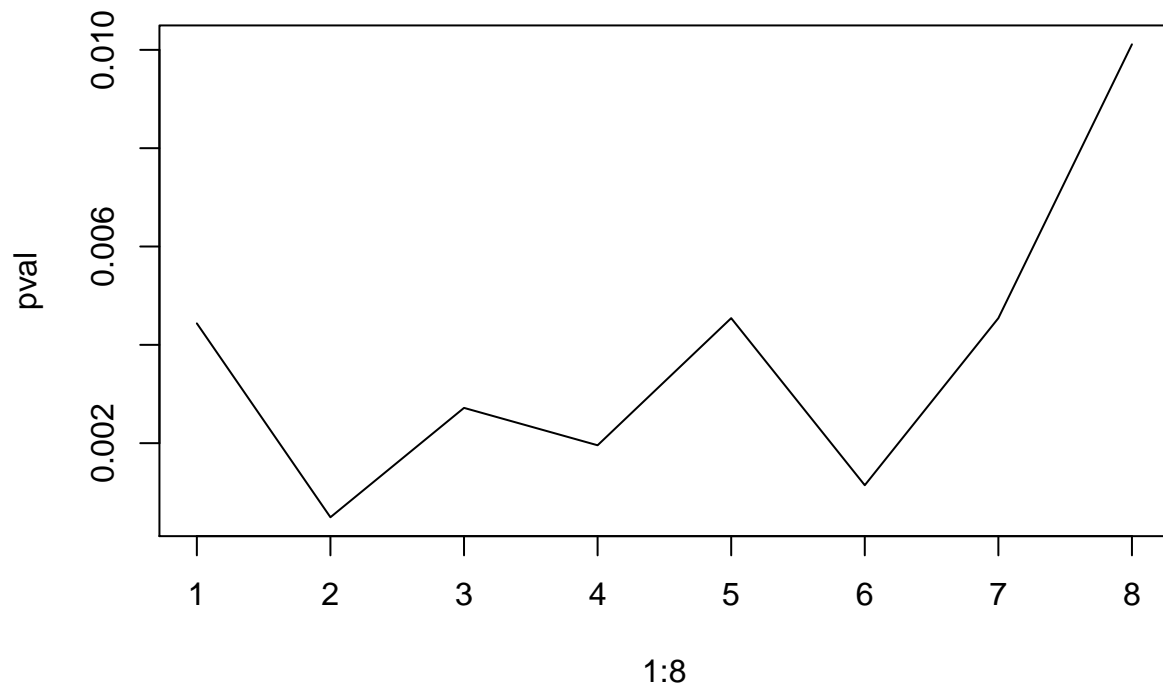
```
pval
```

```
## [1] 0.0044365532 0.0004933992 0.0027173202 0.0019569638 0.0045427413  
## [6] 0.0011444224 0.0045427413 0.0101118390
```

```
plot(1:8,beta,"l")
```



```
plot(1:8,pval,"l")
```



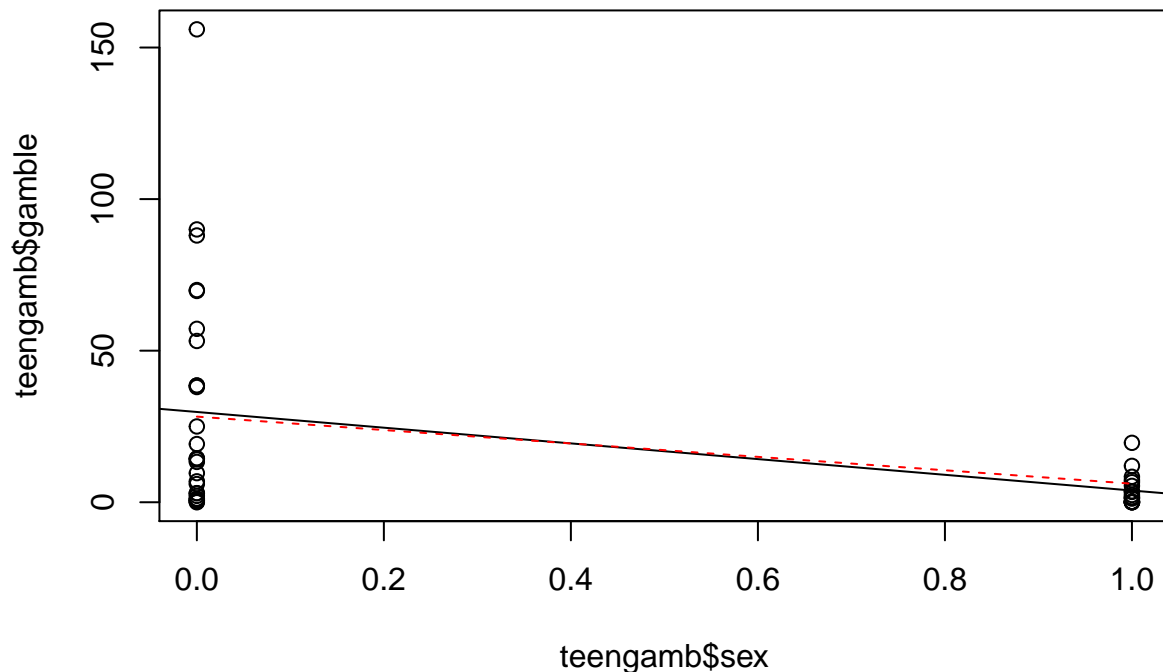
Additional Analysis

Also comparing two cases, first when only sex is included as predictor and second when all the variables are included but for prediction all the variables other than sex is kept constant. It is found that the two model lines are very close to each other supporting the assertion that the effect of sex on response is stable.

```
plot(teengamb$gamble~teengamb$sex)
abline(lmod1)
colMeans(teengamb)
```

```
##      sex      status      income      verbal      gamble
## 0.4042553 45.2340426  4.6419149  6.6595745 19.3010638
```

```
pred_y <-
predict(lmod8,data.frame(sex=teengamb$sex,status=45.2340,income=4.6419,verbal=6.6596))
i <- order(teengamb$sex)
lines(teengamb$sex[i],pred_y[i],lty=2,col=2)
```



Chapter 6 Exercise 1

First the dataset `sat` is loaded from the `faraway` package. The library `faraway` is also loaded because `halfnorm` function required in part(c) is included in this package. Then the first column containing row names is named `states` to make our calculations easier. Then a linear model is fitted for `total` as a response and `expend`, `salary`, `ratio` and `takers` as predictors. This is assigned to a variable `lmod` which is reference in all parts of questions.

Part (a): The fitted values are plotted against residuals. The plot shows that relationship is linear and the constant variance assumption doesn't hold.

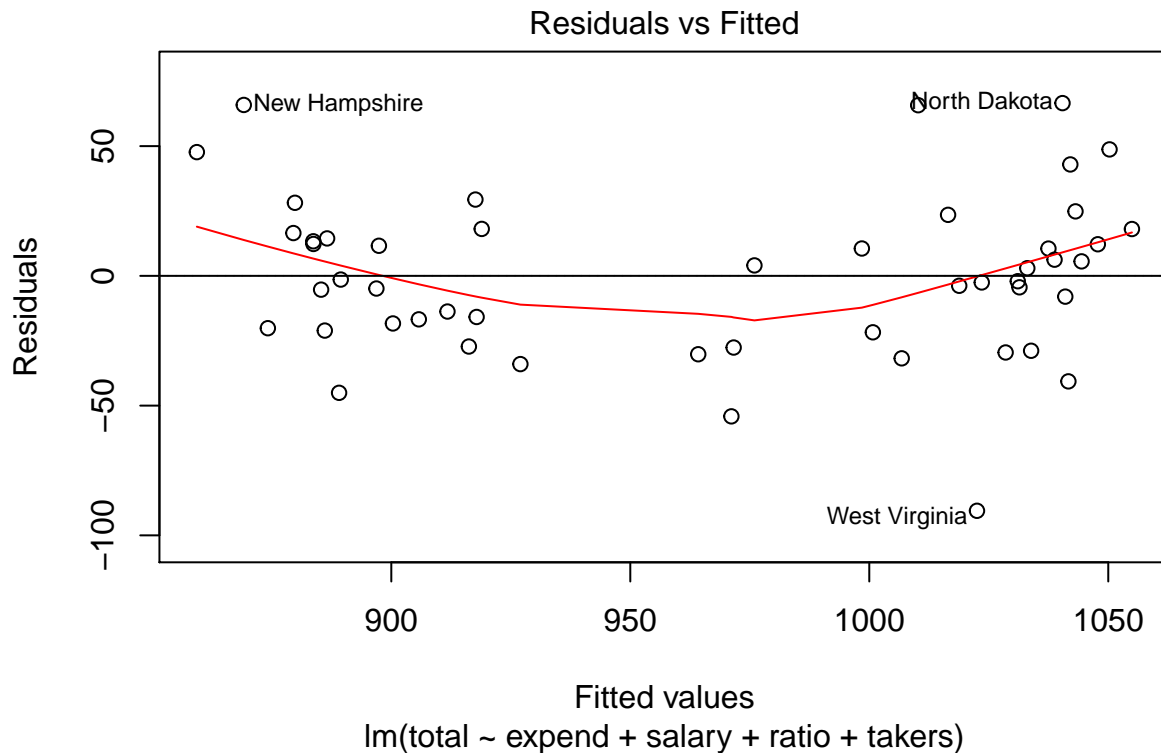
```
data(sat, package='faraway')
library(faraway)

## Warning: package 'faraway' was built under R version 3.4.4
states<-row.names(sat)
lmod<-lm(total~expend+salary+ratio+takers,sat)
summary(lmod)

##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -90.531 -20.855  -1.746   15.979   66.571
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784 < 2e-16 ***
## expend       4.4626     10.5465   0.423  0.674
## salary       1.6379      2.3872   0.686  0.496
## ratio       -3.6242      3.2154  -1.127  0.266
## takers      -2.9045      0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
plot(lmod,which=c(1,1))
abline(h=0)
```



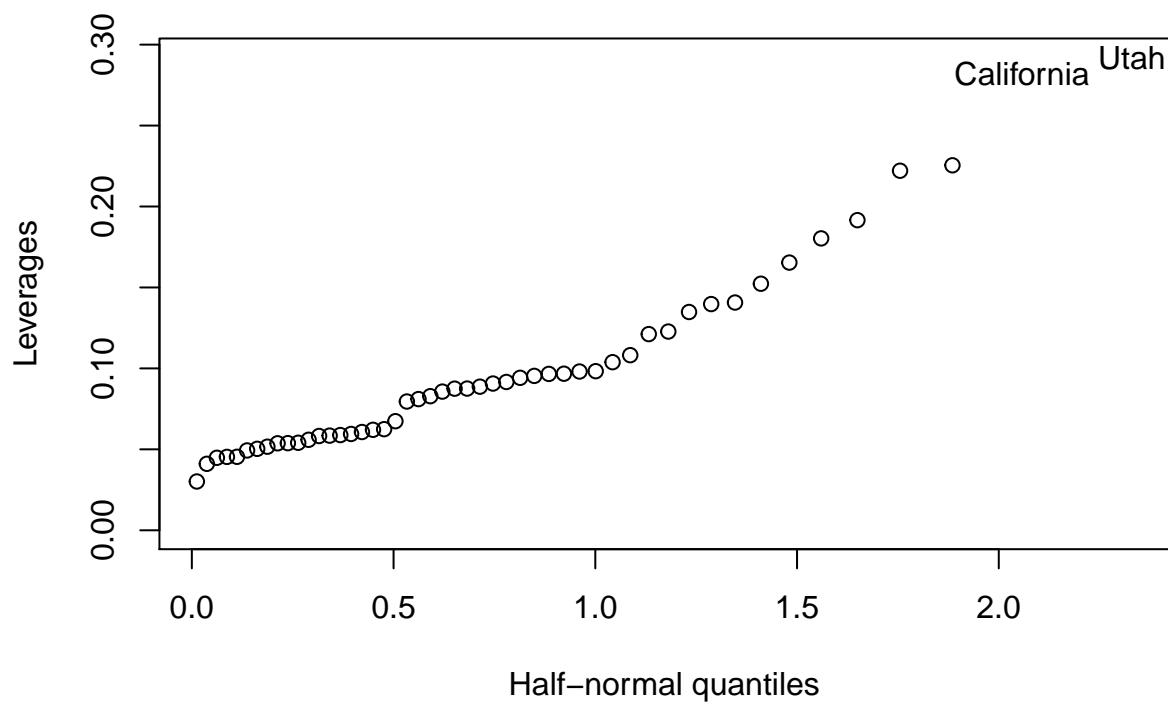
Part (b): Shapiro-Wilk test is used to check the normality assumption. The p-value of 0.4304 means that we cannot reject the null hypothesis that the residuals are normal.

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

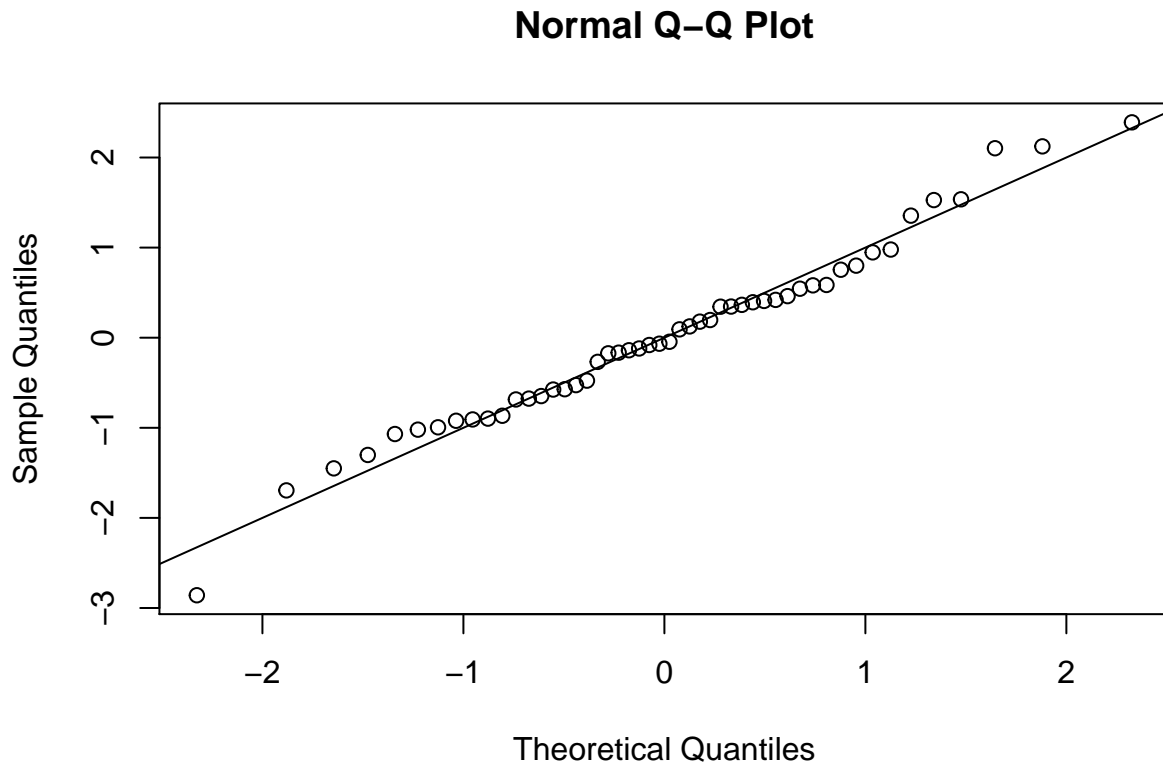
##
## Shapiro-Wilk normality test
##
## data:  residuals(lmod)
## W = 0.97691, p-value = 0.4304
```

Part (c): The half normal plot and the qqnorm plot of standardised residuals is plotted to find out the points with leverage. It is evident from the plots that there are two states Utah and California which have highest leverage although the standardised residuals values are around 2 which is not abnormally high.

```
hatv <- hatvalues(lmod)
halfnorm(hatv, labs=states, ylab="Leverages")
```

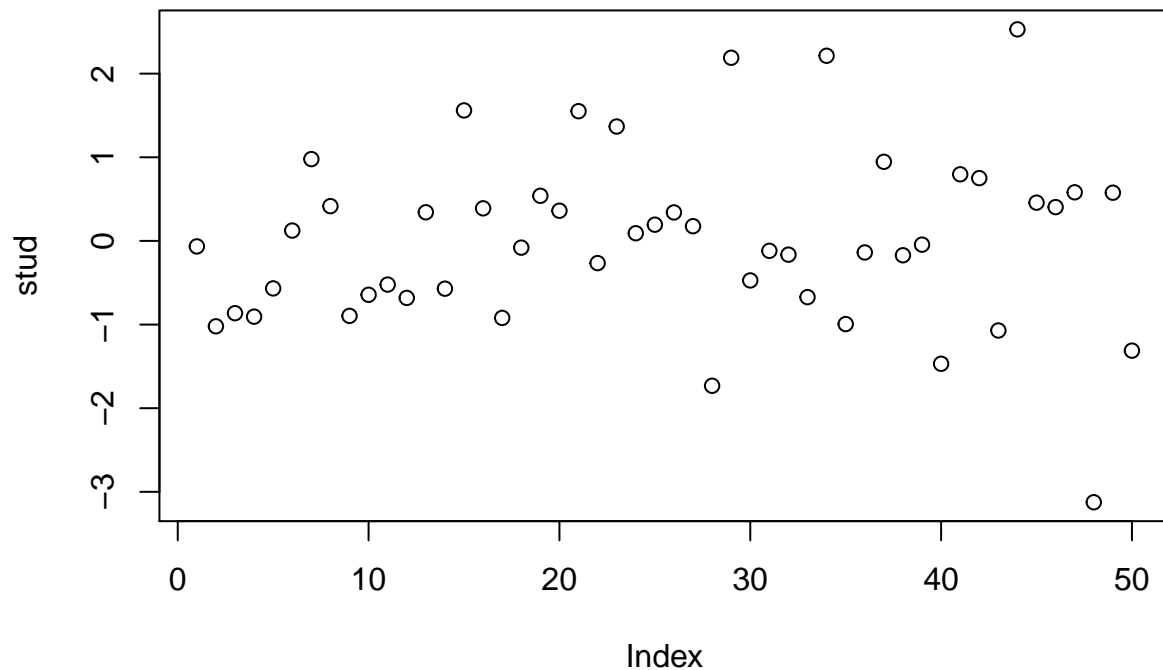


```
qqnorm(rstandard(lmod))
abline(0,1)
```



Part (d): In this part the studentised residuals for the sat data is calculated and the largest value is picked which comes to be -3.12 as from the plot of the studentised residuals it can be seen that there is no masking of outliers so if the maximum value is not a outlier, there will not be another one. The Benferroni correction critical value is computed and it comes to be -3.52 whose absolute value is greater than the absolute value of maximum of the studentised residual which means that the largest value is not an outlier.

```
stud<-rstudent(lmod)
plot(stud)
```



```
stud[which.max(abs(stud))]
```

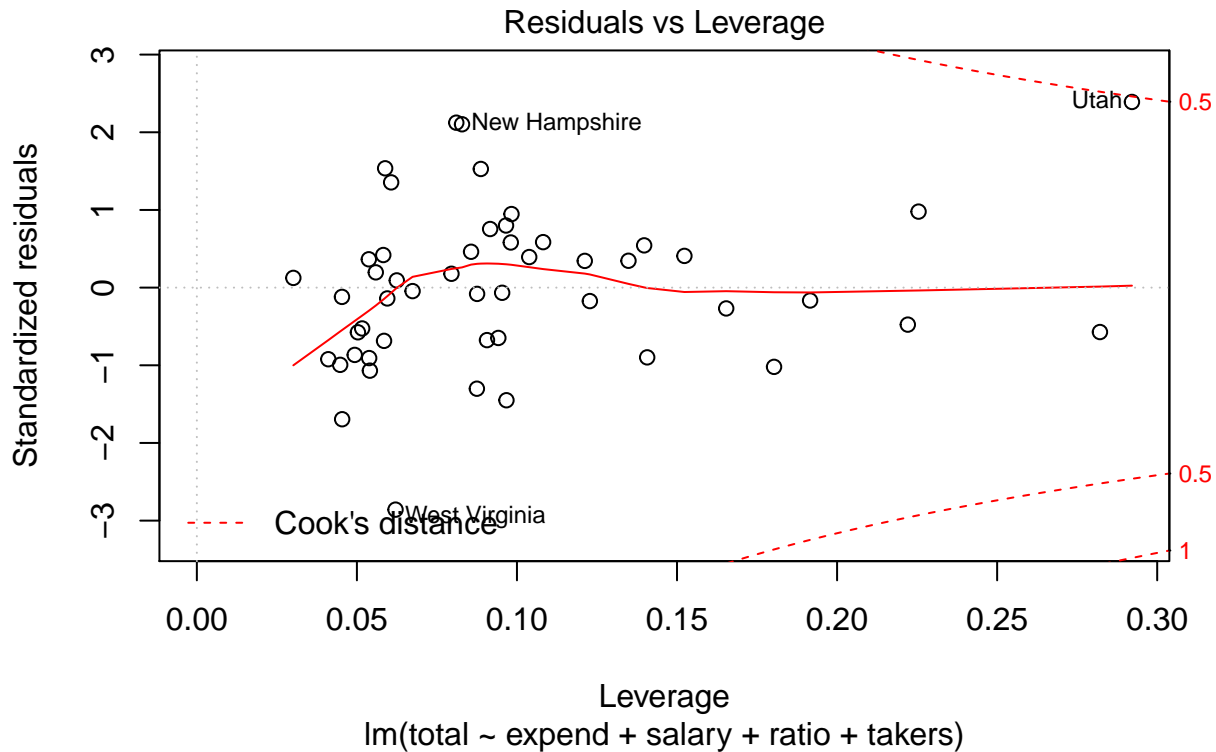
```
## West Virginia
##      -3.124428
```

```
qt(.05/(50*2),lmod$df)
```

```
## [1] -3.520251
```

Part(e): Plotting `lmod` which models response against all predictors was plotted. From the fourth plot: Residuals vs Leverage, it is found that there are 3 suspect influential points: New Hampshire, West Virginia and Utah. We check these points. For Utah, we find that ratio becomes a significant predictor which was not the case earlier, So, it is most certainly an influential point. For cases of New Hampshire and West Virginia, the coefficients do not change much and the R-squared value remains almost same, so these two states should not be considered as influential points.

```
plot(lmod,which=c(5,5))
```

```
lmodel1<-lm(total~expend+salary+ratio+takers,sat,subset=(states!="Utah"))
summary(lmodel1)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,
##     subset = (states != "Utah"))
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
##	-92.118	-18.402	1.808	14.890	67.669

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	1093.8460	53.4226	20.475	<2e-16 ***
## expend	-0.9427	10.1922	-0.092	0.927
## salary	3.0964	2.3283	1.330	0.190
## ratio	-7.6391	3.4279	-2.229	0.031 *
## takers	-2.9308	0.2188	-13.397	<2e-16 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.9 on 44 degrees of freedom
## Multiple R-squared:  0.8396, Adjusted R-squared:  0.825
## F-statistic: 57.58 on 4 and 44 DF,  p-value: < 2.2e-16
```

```
lmode2<-lm(total~expend+salary+ratio+takers,sat,subset=(states!="New Hampshire"))
summary(lmode2)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,
##     subset = (states != "New Hampshire"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -90.611 -15.308  -1.274   19.097   67.054
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1031.3254    51.2102  20.139  <2e-16 ***
## expend         6.2629    10.1613   0.616   0.541
## salary         1.5893     2.2926   0.693   0.492
## ratio        -3.1407     3.0957  -1.015   0.316
## takers        -3.0146     0.2277 -13.239  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.4 on 44 degrees of freedom
## Multiple R-squared:  0.8412, Adjusted R-squared:  0.8268
## F-statistic: 58.29 on 4 and 44 DF,  p-value: < 2.2e-16
```

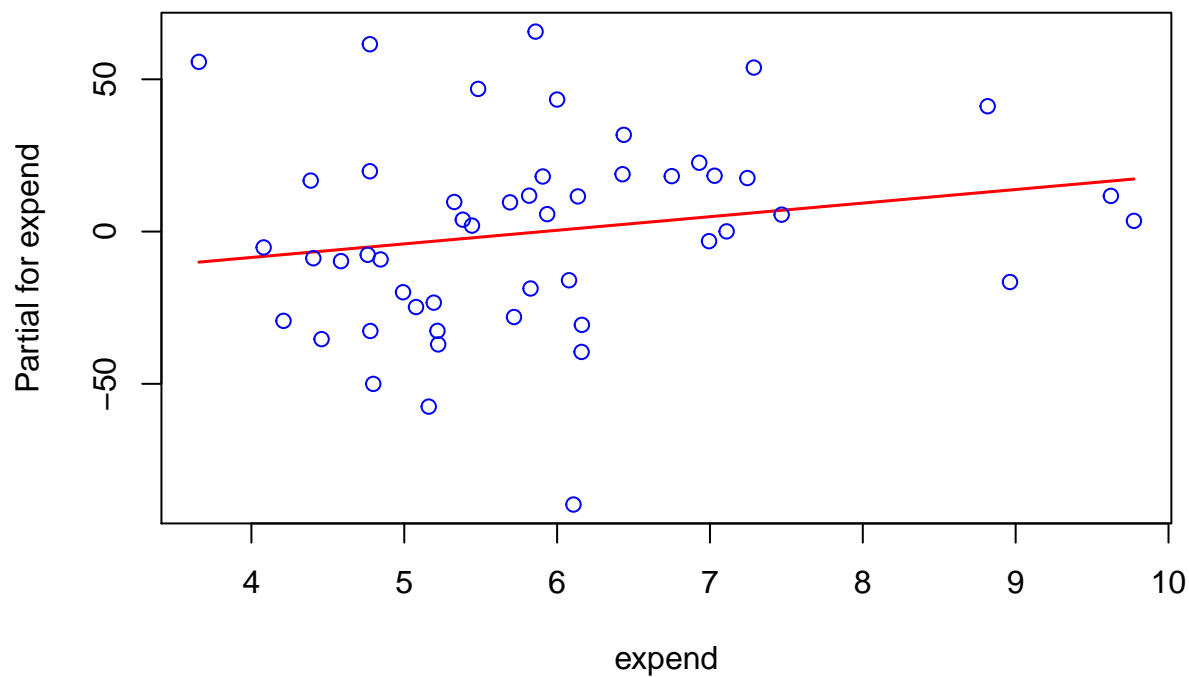
```
lmode3<-lm(total~expend+salary+ratio+takers,sat,subset=(states!="West Virginia"))
summary(lmode3)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,
##     subset = (states != "West Virginia"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.692 -19.589  -2.957   12.737   67.414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1057.6775    48.5148  21.801  <2e-16 ***
## expend         7.3593     9.6933   0.759   0.452
## salary         1.0875     2.1912   0.496   0.622
## ratio        -3.9397     2.9435  -1.338   0.188
## takers        -2.9724     0.2127 -13.975  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.92 on 44 degrees of freedom
## Multiple R-squared:  0.8558, Adjusted R-squared:  0.8427
## F-statistic: 65.28 on 4 and 44 DF,  p-value: < 2.2e-16
```

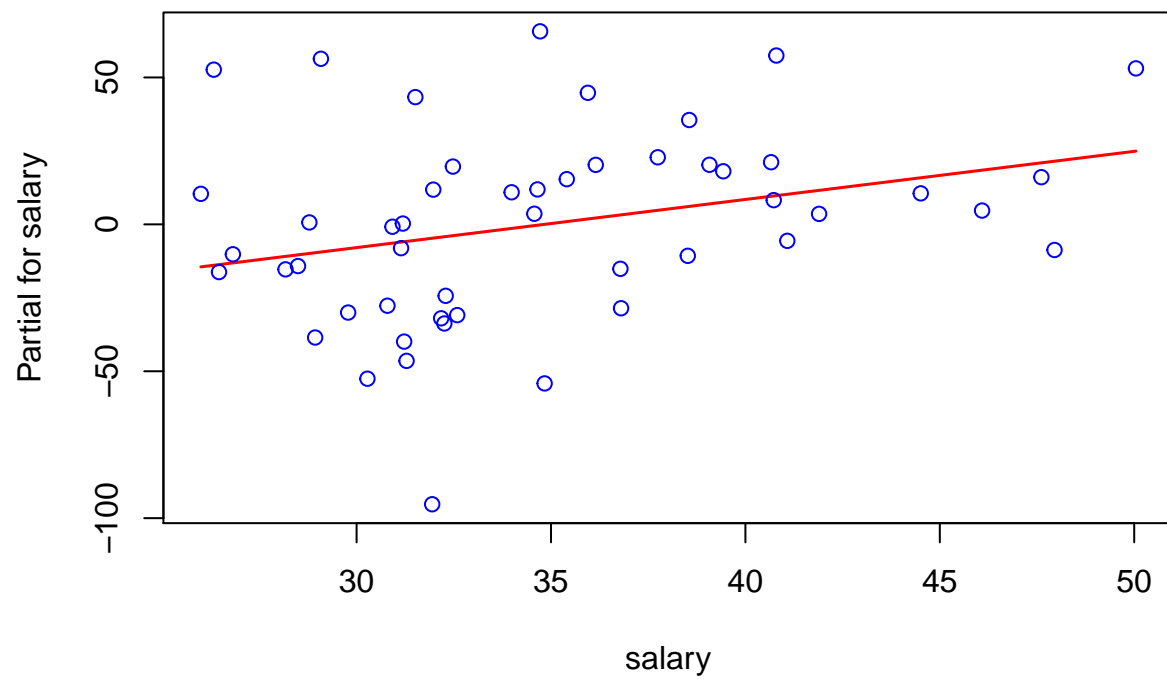
Part(f): Plotting partial residuals for every predictor, it is found that for the case of predictor: “taker”, we see formation of two groups. So we investigate by forming two subsets, one for takers< 40 and one for takers>40. We can see from the results that for takers<40, takers is the only significant predictor and for takers >40,

none of the predictors are significant. We can use this result to simplify the structure of our model.

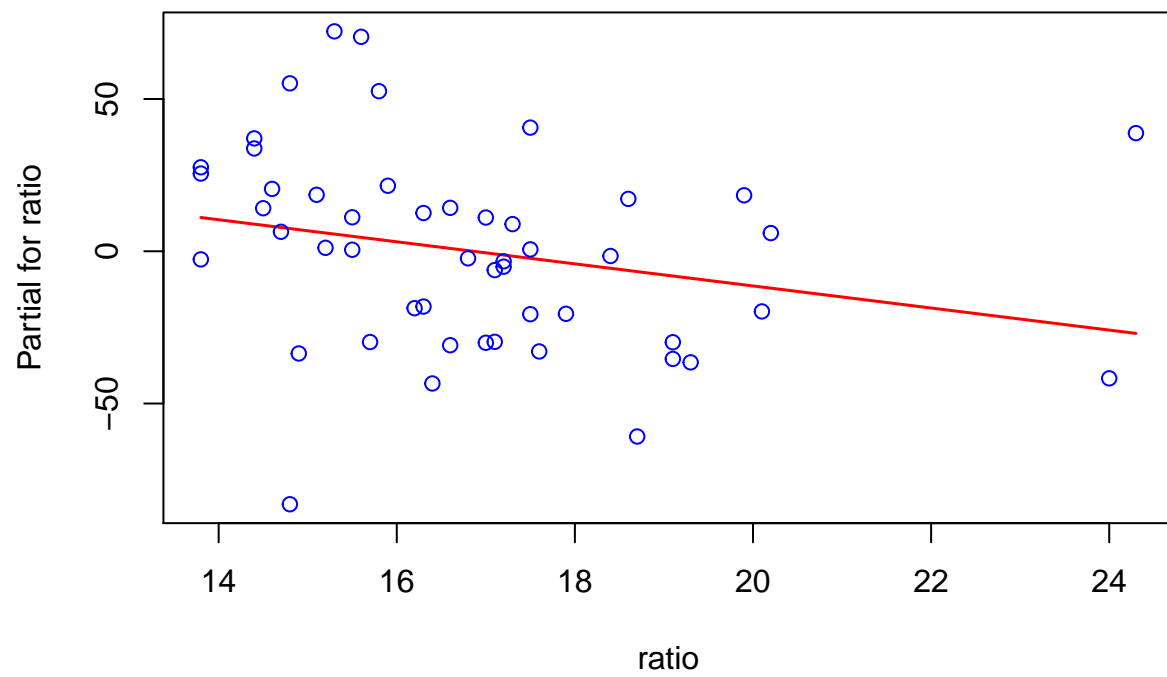
```
lmod<-lm(total~expend+salary+ratio+takers,sat)
termplot(lmod,partial.resid = TRUE,terms=1,col.res="blue")
```



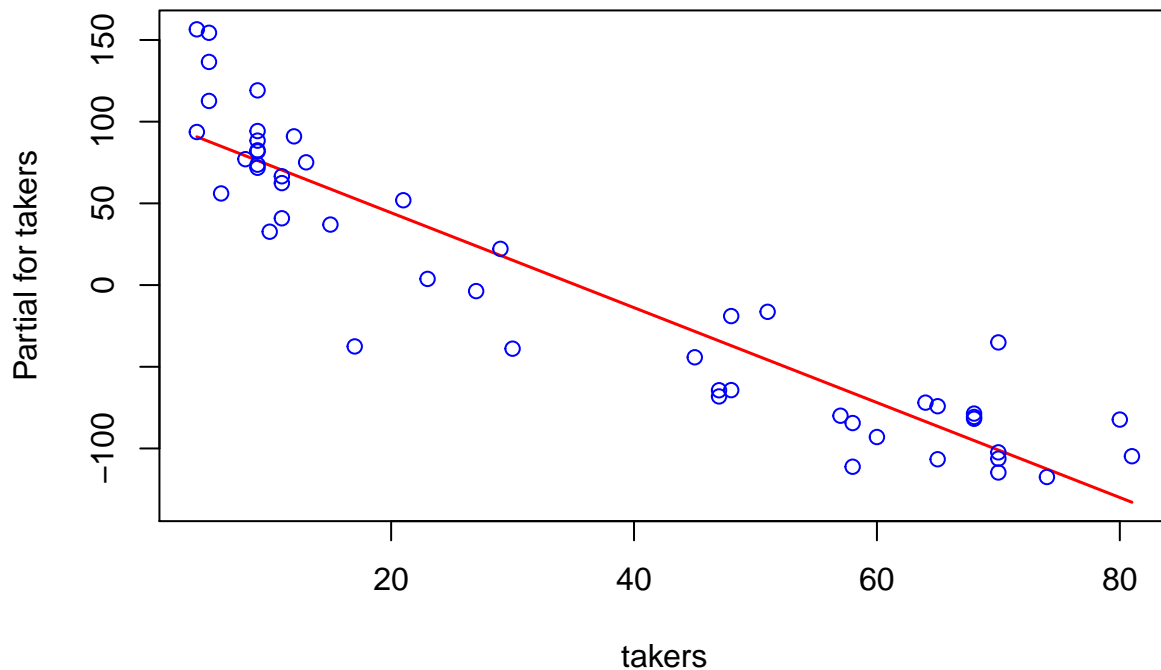
```
termplot(lmod,partial.resid = TRUE,terms=2,col.res="blue")
```



```
termplot(lmod,partial.resid = TRUE,terms=3,col.res="blue")
```



```
termplot(lmod,partial.resid = TRUE,terms=4,col.res="blue")
```



```
lmodf1<-lm(total~expend+salary+ratio+takers,sat,subset=(takers<40))
lmodf2<-lm(total~expend+salary+ratio+takers,sat,subset=(takers>40))
```

```
summary(lmodf1)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,
##     subset = (takers < 40))
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-69.155	-16.672	1.747	15.654	54.956

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	993.7178	84.5010	11.760	5.86e-11 ***
expend	7.7581	16.4329	0.472	0.641
salary	1.0293	3.3058	0.311	0.758
ratio	1.4251	4.6111	0.309	0.760
takers	-5.5242	0.8706	-6.345	2.19e-06 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.95 on 22 degrees of freedom
## Multiple R-squared:  0.656, Adjusted R-squared:  0.5935
```

```
## F-statistic: 10.49 on 4 and 22 DF,  p-value: 6.549e-05
summary(lmodf2)

##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,
##     subset = (takers > 40))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -38.434 -15.938  -2.509   11.444   50.289
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  801.4329   105.6773    7.584 5.2e-07 ***
## expend       11.1444    10.8359    1.028  0.317
## salary      -0.6354     2.7190   -0.234  0.818
## ratio         3.9147     4.8627    0.805  0.431
## takers      -0.3003     0.8869   -0.339  0.739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.74 on 18 degrees of freedom
## Multiple R-squared:  0.2627, Adjusted R-squared:  0.09882
## F-statistic: 1.603 on 4 and 18 DF,  p-value: 0.2167
```

Chapter 6 Exercise 3

First the dataset prostate is loaded from the faraway package. The library faraway is also loaded because halfnorm function required in part(c) is included in this package. Then the first column containing row indices is named indices to make our calculations easier. Then a linear model is fitted for total as a response and lcaivol, lweight, age, lbph, svi, scp, gleason and pgg45 as predictors. This is assigned to a variable lmod which is reference in all parts of questions.

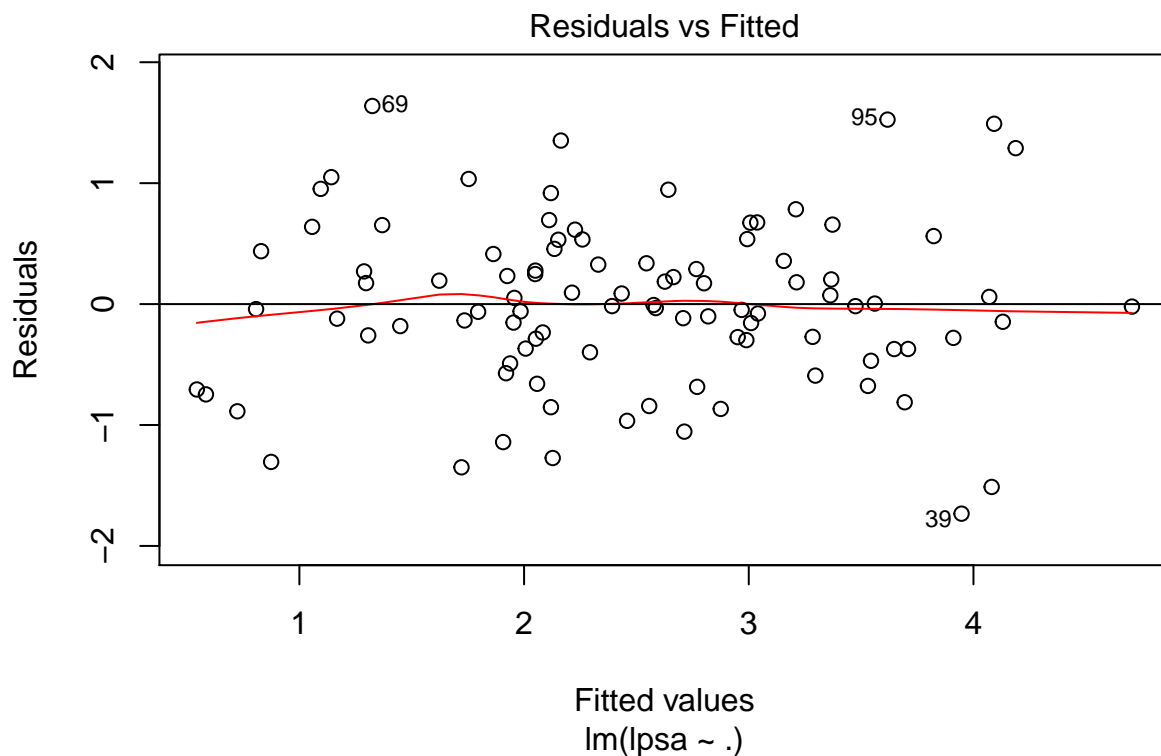
Part (a): The fitted values are plotted against residuals. The plot shows that the constance variance assumption holds.

```
data(prostate,package='faraway')
library(faraway)
indices<-row.names(prostate)
lmod<-lm(lpsa~.,prostate)
summary(lmod)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
```

```
## lcavol      0.587022   0.087920   6.677 2.11e-09 ***
## lweight     0.454467   0.170012   2.673 0.00896 **
## age        -0.019637   0.011173  -1.758 0.08229 .
## lbph       0.107054   0.058449   1.832 0.07040 .
## svi        0.766157   0.244309   3.136 0.00233 **
## lcp       -0.105474   0.091013  -1.159 0.24964
## gleason    0.045142   0.157465   0.287 0.77503
## pgg45      0.004525   0.004421   1.024 0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
plot(lmod,which=c(1,1))
abline(h=0)
```



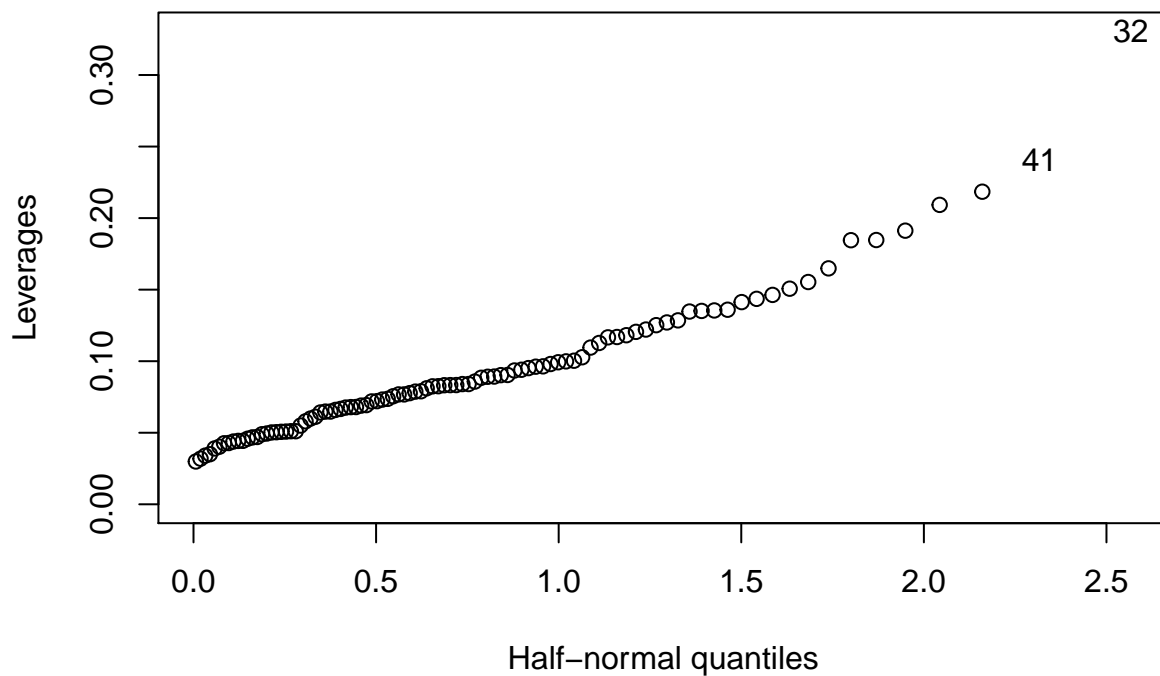
Part (b): Shapiro-Wilk test is used to check the normality assumption. The p-value of 0.7721 means that we cannot reject the null hypothesis that the residuals are normal.

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

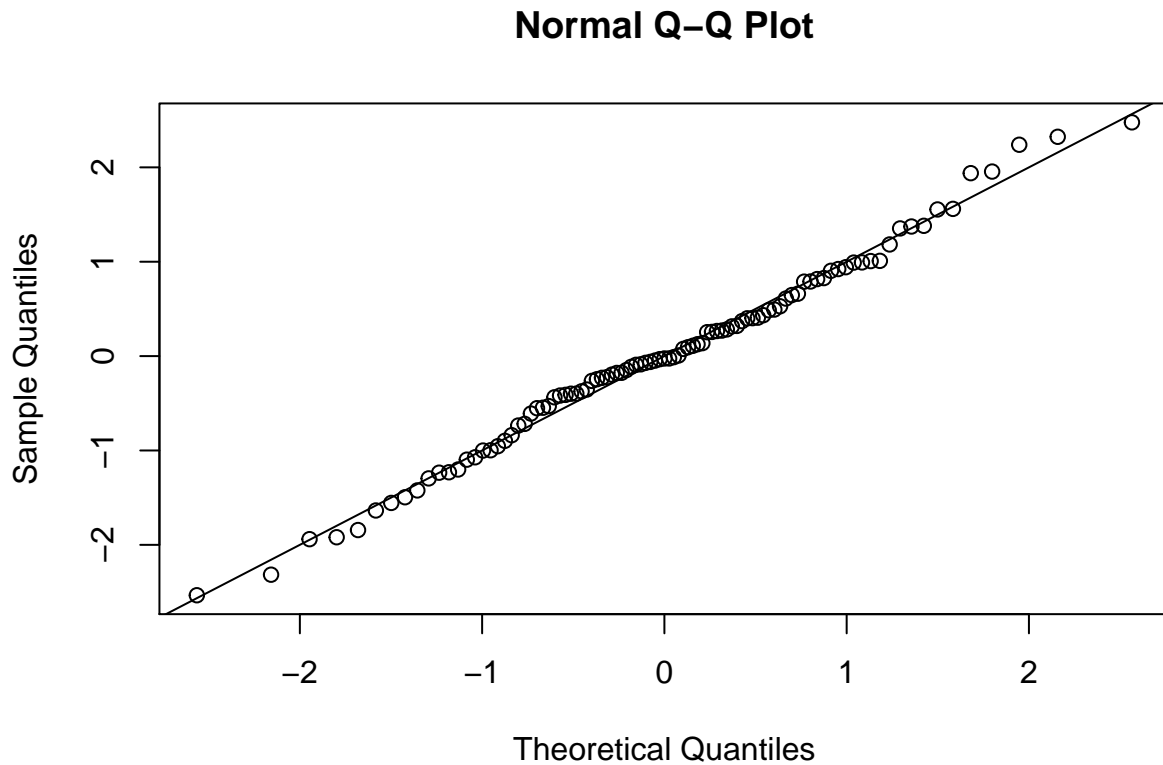
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmod)
## W = 0.99113, p-value = 0.7721
```


Part (c): The half normal plot and the qqnorm plot of standardised residuals is plotted to find out the points with leverage. It is evident from the plots that there are two states Utah and California which have highest leverage although the standardised residuals values are around 2 which is not abnormally high.

```
hatv <- hatvalues(lmod)
halfnorm(hatv, labs=indices, ylab="Leverages")
```

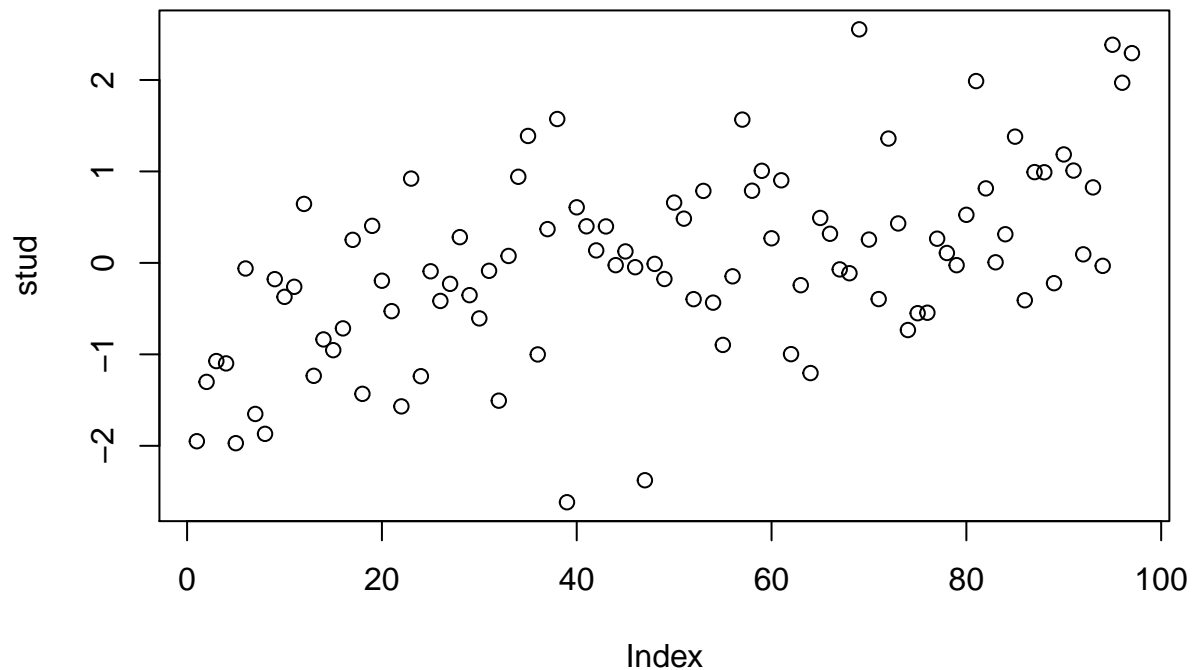


```
qqnorm(rstandard(lmod))
abline(0,1)
```



Part (d): In this part the studentised residuals for the sat data is calculated and the largest value is picked which comes to be -2.62 as from the plot of the studentised residuals it can be seen that there is no masking of outliers so if the maximum value is not a outlier, there will not be another one. The Benferroni correction critical value is computed and it comes to be -3.60 whose absolute value is greater than the absolute value of maximum of the studentised residual which means that the largest value is not an outlier.

```
stud<-rstudent(lmod)
plot(stud)
```



```
stud[which.max(abs(stud))]
```

```
##      39
## -2.61698
```

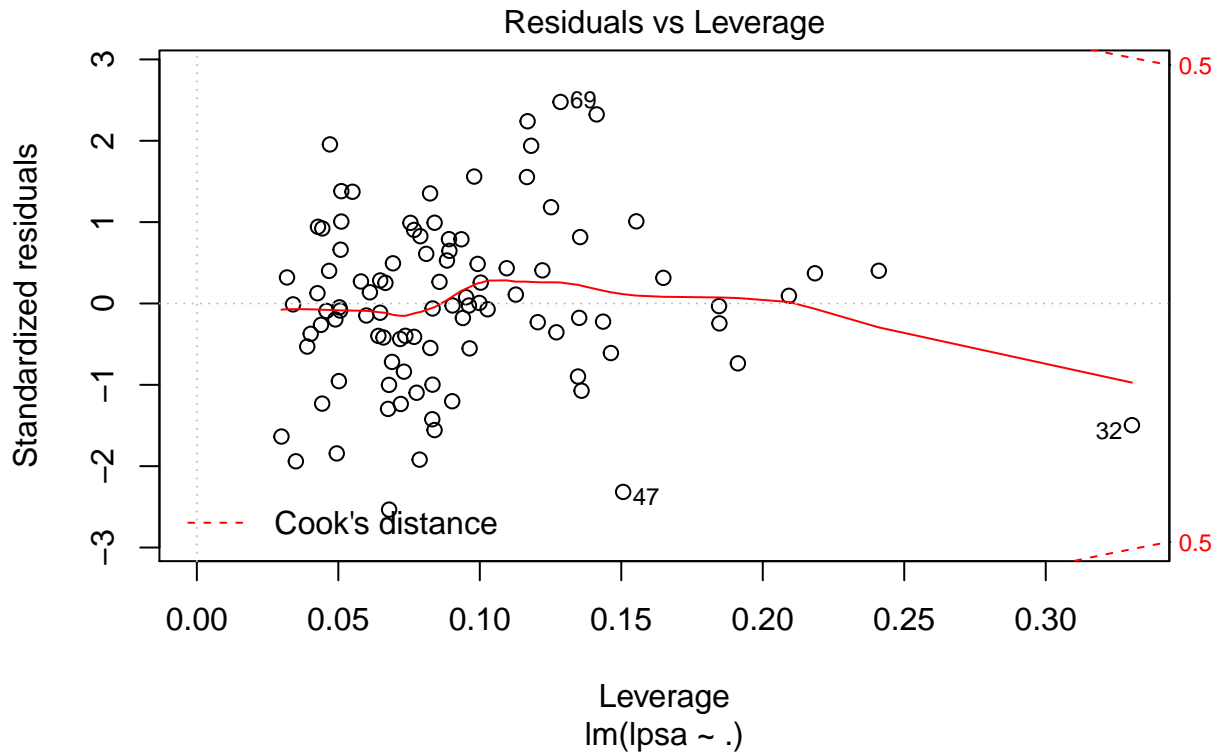
```
qt(.05/(nrow(prostate)*2),lmod$df)
```

```
## [1] -3.605841
```

Part(e): Plotting `lmod` which models response against all predictors was plotted. From the fourth plot: Residuals vs Leverage, it is found that there are 3 suspect influential points: 69, 47 and 32. We check these points. For 69, we find that `age` and `lweight` have become significant predictors which were not so in original models implying that 69 is an influential observation. For 47, we find that `lweight` has become significant predictor which was not so in original models implying that 47 is an influential observation.

For 32, we find that there is no change in significance and coefficients. Hence this observation is not influential.

```
plot(lmod,which=c(5,5))
```



```
lmodel1<-lm(lpsa~.,prostate,subset=(indices!=69))
summary(lmodel1)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (indices !=
##      69))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73872 -0.33997 -0.00536  0.41060  1.48914
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.156567   1.271936   0.909  0.36571
## lcavol       0.628361   0.086809   7.238 1.7e-10 ***
## lweight      0.350500   0.169870   2.063  0.04206 *
## age         -0.023361   0.010936  -2.136  0.03547 *
## lbph        0.141597   0.058289   2.429  0.01719 *
## svi         0.773688   0.237007   3.264  0.00157 **
## lcp        -0.116974   0.088401  -1.323  0.18923
## gleason      0.051287   0.152765   0.336  0.73789
## pgg45       0.004792   0.004290   1.117  0.26701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6872 on 87 degrees of freedom
## Multiple R-squared:  0.6782, Adjusted R-squared:  0.6486
## F-statistic: 22.92 on 8 and 87 DF,  p-value: < 2.2e-16

lmode2<-lm(lpsa~.,prostate,subset=(indices!=47))
summary(lmode2)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (indices !=
##      47))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82761 -0.33555 -0.00808  0.42191  1.65172
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.048804   1.290142   0.038 0.969911
## lcavol       0.565850   0.086148   6.568 3.57e-09 ***
## lweight      0.457651   0.165698   2.762 0.007009 **
## age         -0.016845   0.010952  -1.538 0.127655
## lbph         0.116665   0.057107   2.043 0.044087 *
## svi          0.826463   0.239449   3.452 0.000863 ***
## lcp         -0.097979   0.088757  -1.104 0.272681
## gleason      0.114229   0.156192   0.731 0.466540
## pgg45        0.004457   0.004309   1.034 0.303832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6904 on 87 degrees of freedom
## Multiple R-squared:  0.6758, Adjusted R-squared:  0.646
## F-statistic: 22.67 on 8 and 87 DF,  p-value: < 2.2e-16

lmode3<-lm(lpsa~.,prostate,subset=(indices!=32))
summary(lmode3)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (indices !=
##      32))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76425 -0.35383 -0.03029  0.40930  1.55613
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.171863   1.328822   0.129 0.89739
## lcavol       0.565333   0.088472   6.390 7.93e-09 ***
## lweight      0.621663   0.202017   3.077 0.00279 **
## age         -0.021271   0.011146  -1.908 0.05963 .
## lbph         0.095590   0.058529   1.633 0.10604
## svi          0.760423   0.242596   3.135 0.00235 **
## lcp         -0.105987   0.090365  -1.173 0.24404
```

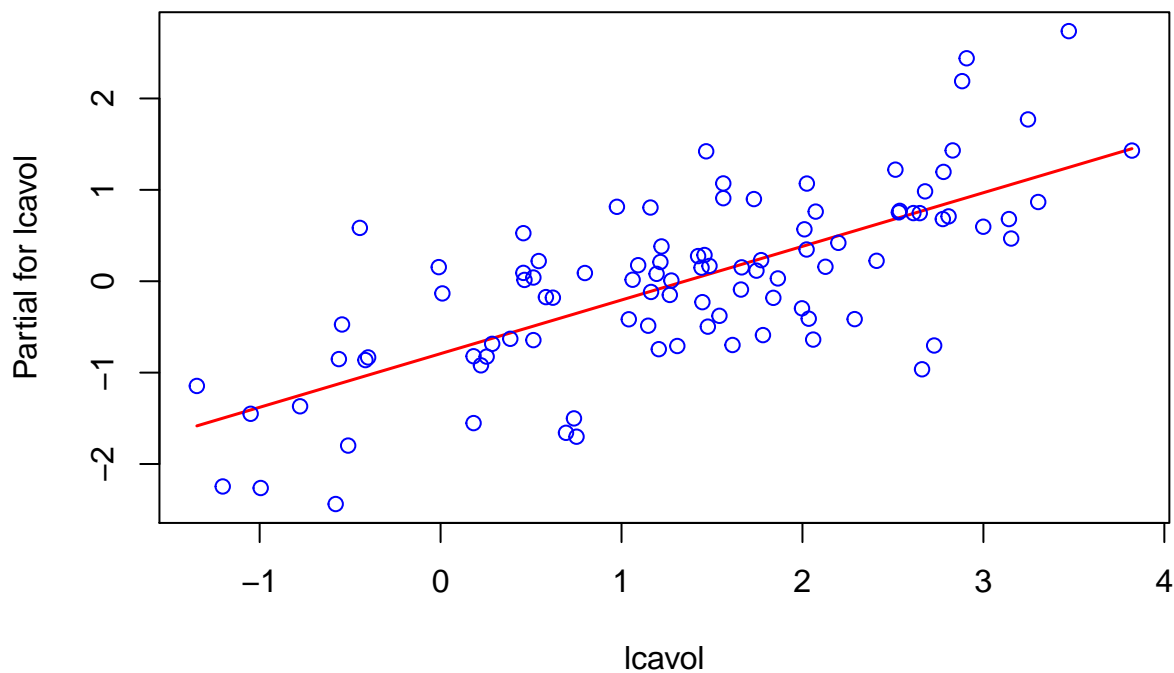
```
## gleason      0.050688  0.156384  0.324  0.74662
## pgg45       0.004468  0.004390  1.018  0.31155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7034 on 87 degrees of freedom
## Multiple R-squared:  0.6629, Adjusted R-squared:  0.6319
## F-statistic: 21.39 on 8 and 87 DF,  p-value: < 2.2e-16
```

Part(f): Plotting partial residuals for every predictor, it is found that for the case of predictors: lbph, svi, lcp, gleason and pgg45 there are groups. These groups are due to discrete nature of these variables. It holds true for svi because it is categorical. For svi=0, model is similar. For svi==1, none of the variables are significant.

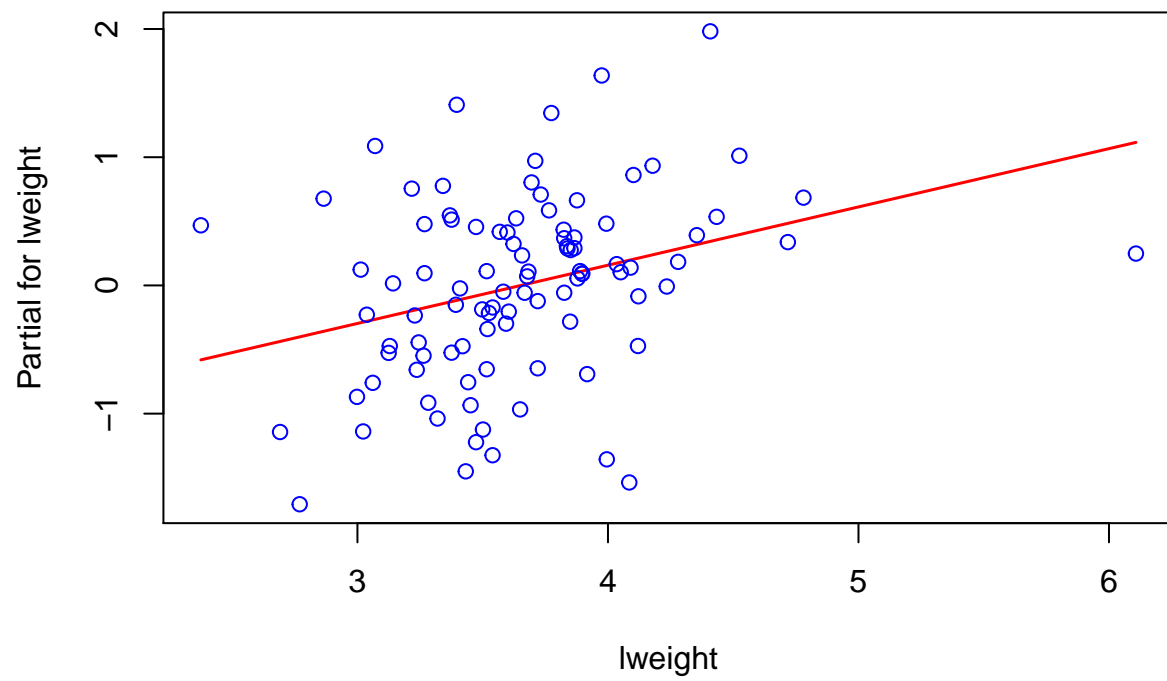
For lbph<-1, we have similar model but for lbph>=-1, only lcavol is significant.

For lcp<-1, lcavol and lweight are significant. For lcp>-1, lcavol and svi are significant.

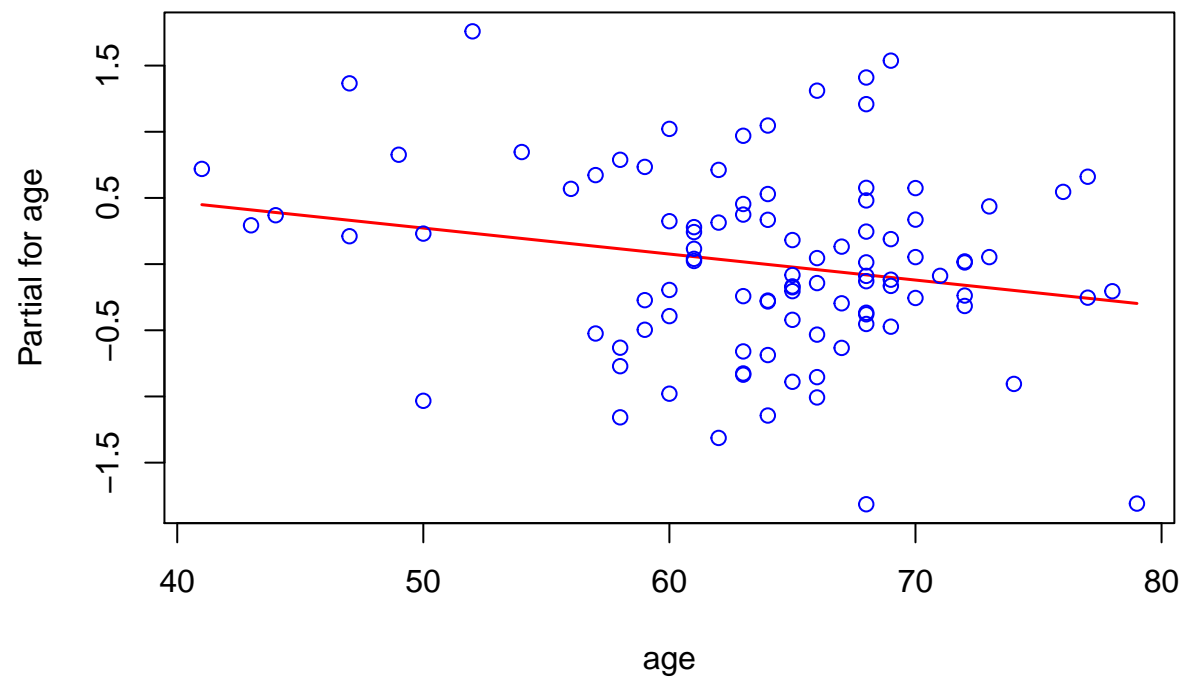
```
termplot(lmod,partial.resid = TRUE,terms=1,col.res="blue")
```



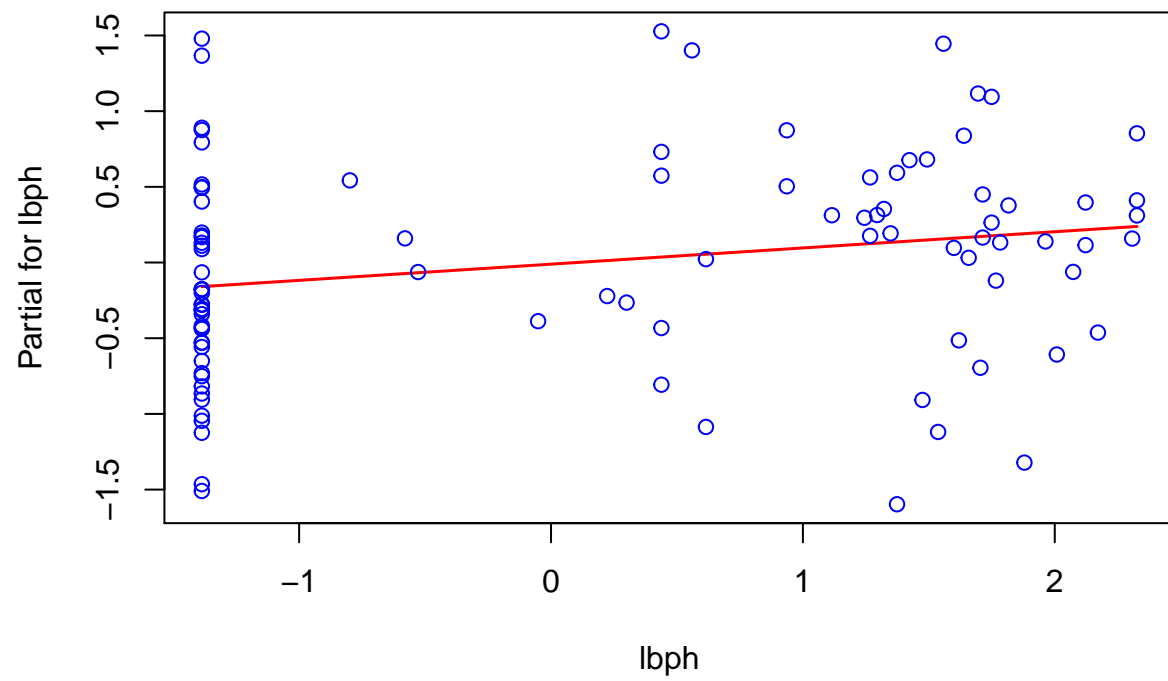
```
termplot(lmod,partial.resid = TRUE,terms=2,col.res="blue")
```



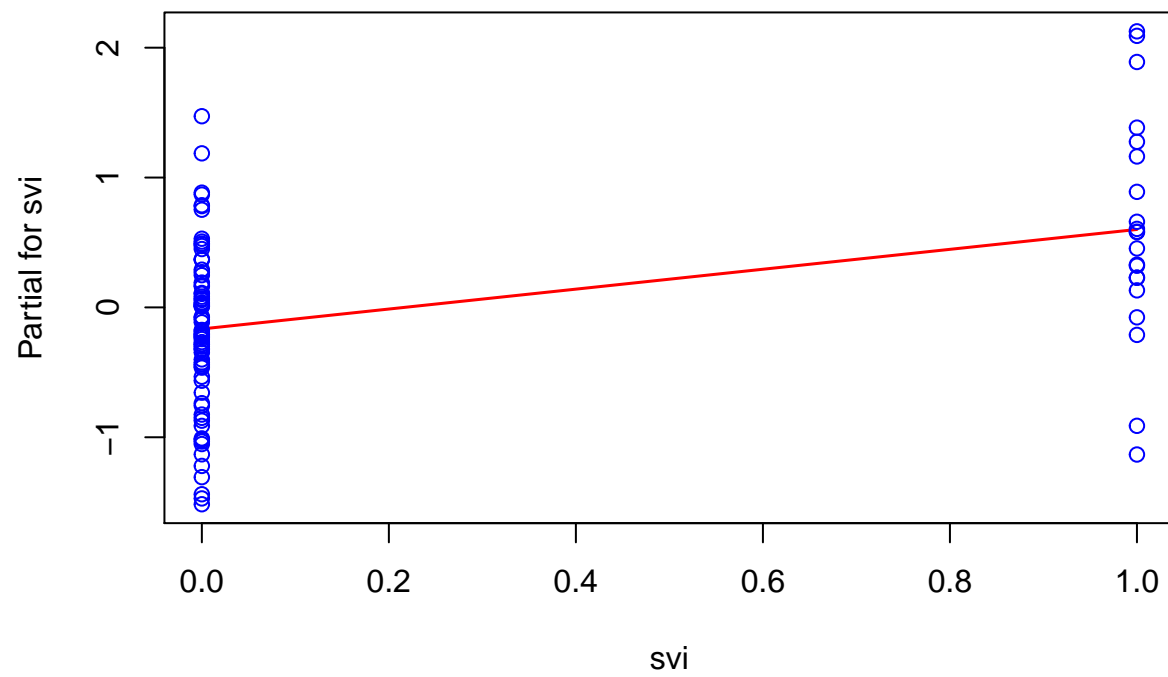
```
termplot(lmod,partial.resid = TRUE,terms=3,col.res="blue")
```



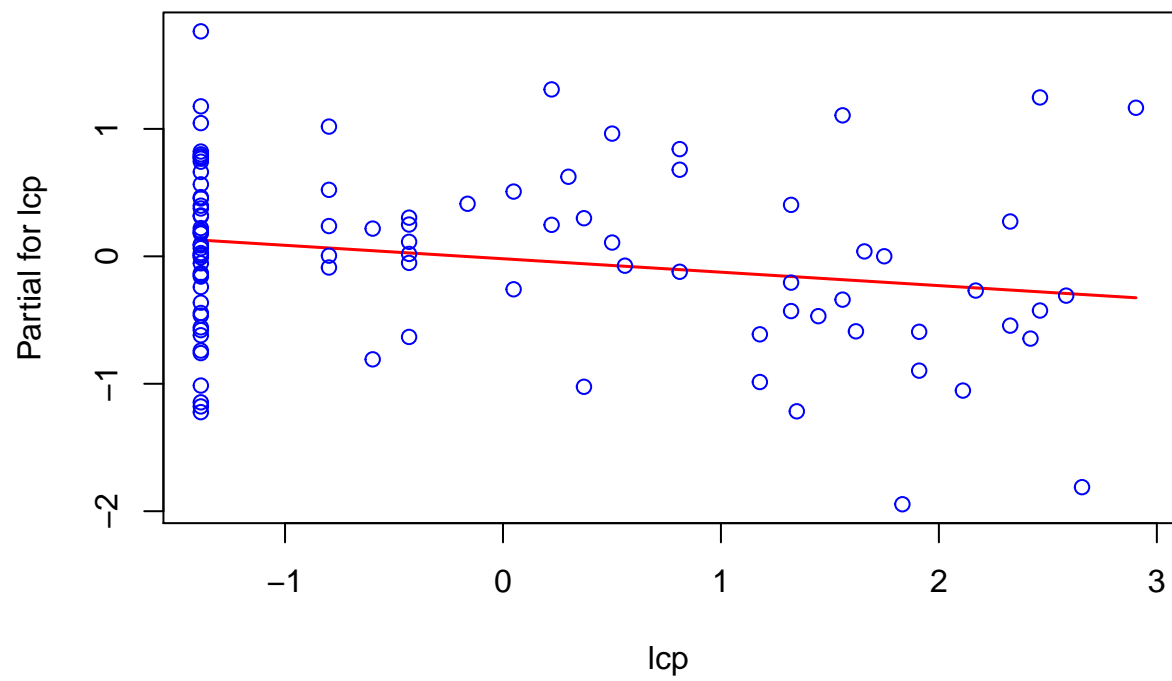
```
termplot(lmod,partial.resid = TRUE,terms=4,col.res="blue")
```

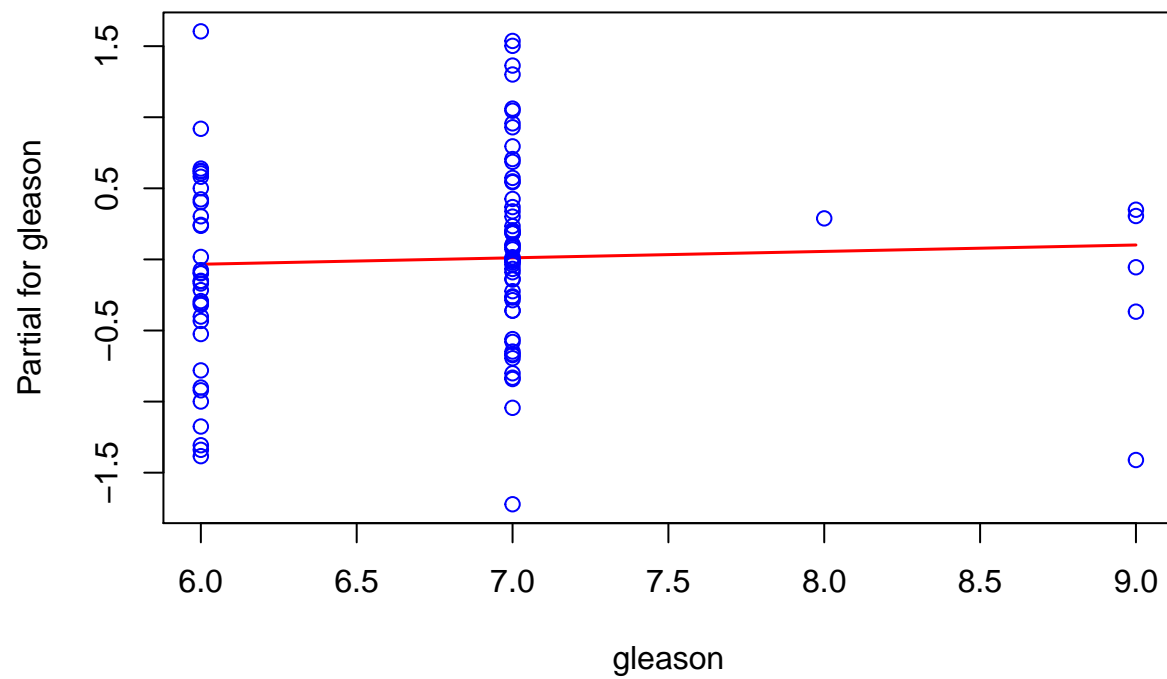
```
termplot(lmod,partial.resid = TRUE,terms=5,col.res="blue")
```



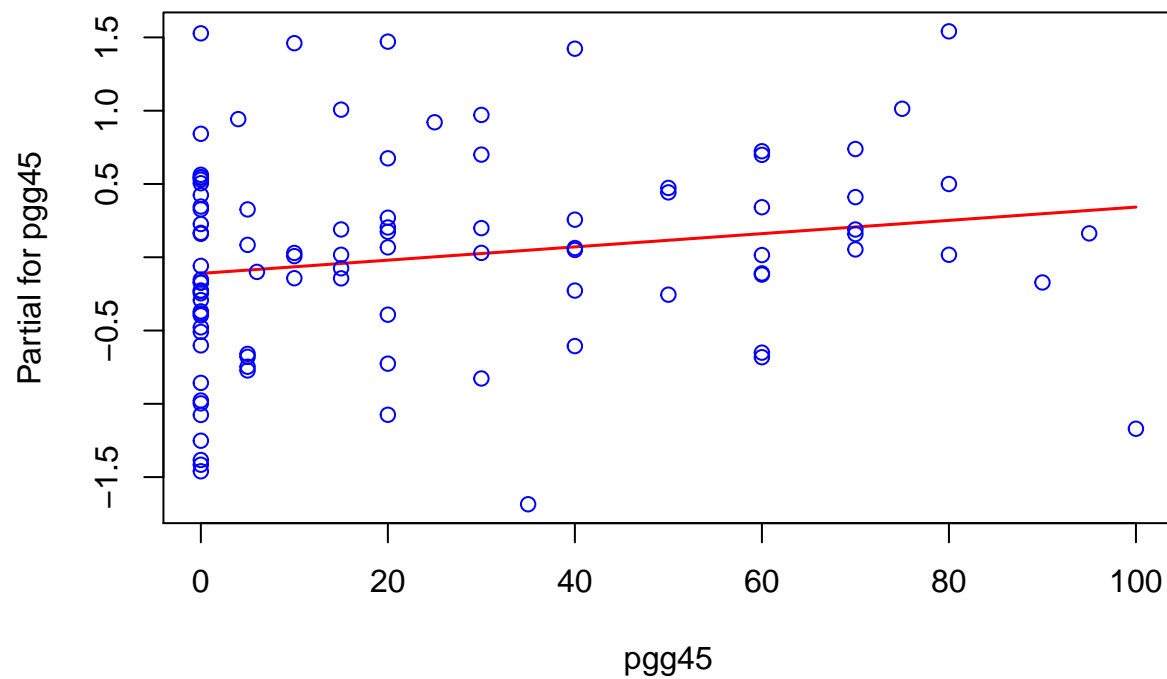
```
termplot(lmod,partial.resid = TRUE,terms=6,col.res="blue")
```



```
termplot(lmod,partial.resid = TRUE,terms=7,col.res="blue")
```



```
termplot(lmod,partial.resid = TRUE,terms=8,col.res="blue")
```



```

lmodf1<-lm(lpsa~.,prostate,subset=(svi==0))

lmodf2<-lm(lpsa~.,prostate,subset=(svi==1))

lmodf3<-lm(lpsa~.,prostate,subset=(lbph<=-1))

lmodf4<-lm(lpsa~.,prostate,subset=(lbph>=-1))

lmodf5<-lm(lpsa~.,prostate,subset=(lcp<=-1))

lmodf6<-lm(lpsa~.,prostate,subset=(lcp>=-1))

summary(lmodf1)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (svi == 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25098 -0.32328 -0.00775  0.37289  1.67838

```

```
##
## Coefficients: (1 not defined because of singularities)
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.109987   1.317289  -0.083  0.93370
## lcavol      0.529821   0.085915   6.167 4.34e-08 ***
## lweight     0.461621   0.167626   2.754  0.00755 **
## age        -0.017300   0.011895  -1.454  0.15044
## lbph       0.146500   0.061087   2.398  0.01922 *
## svi         NA         NA         NA     NA
## lcp        -0.104578   0.096756  -1.081  0.28359
## gleason     0.141251   0.166624   0.848  0.39956
## pgg45       0.005667   0.004769   1.188  0.23886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6468 on 68 degrees of freedom
## Multiple R-squared:  0.5924, Adjusted R-squared:  0.5504
## F-statistic: 14.12 on 7 and 68 DF,  p-value: 3.661e-11
```

```
summary(lmodf2)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (svi == 1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4479 -0.4678 -0.1971  0.2431  1.5776
##
## Coefficients: (1 not defined because of singularities)
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.792941   4.499687   1.510   0.155
## lcavol       0.747157   0.397862   1.878   0.083 .
## lweight     -0.145643   0.688678  -0.211   0.836
## age         -0.019462   0.030472  -0.639   0.534
## lbph        -0.007265   0.170365  -0.043   0.967
## svi          NA         NA         NA     NA
## lcp         -0.203350   0.258871  -0.786   0.446
## gleason     -0.425042   0.442906  -0.960   0.355
## pgg45        0.004484   0.011088   0.404   0.693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9212 on 13 degrees of freedom
## Multiple R-squared:  0.3556, Adjusted R-squared:  0.008541
## F-statistic: 1.025 on 7 and 13 DF,  p-value: 0.459
```

```
summary(lmodf3)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (lbph <= -1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.1813 -0.4483 -0.0918 0.4027 1.4032
##
## Coefficients: (1 not defined because of singularities)
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.730637   2.328484  -1.173  0.24883
## lcavol      0.527630   0.119067   4.431 8.81e-05 ***
## lweight     0.862828   0.304941   2.829  0.00767 **
## age        -0.026725   0.014730  -1.814  0.07820 .
## lbph        NA         NA         NA     NA
## svi         1.142281   0.416713   2.741  0.00958 **
## lcp        -0.128127   0.149317  -0.858  0.39668
## gleason     0.400806   0.270570   1.481  0.14746
## pgg45      -0.002948   0.007539  -0.391  0.69818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7051 on 35 degrees of freedom
## Multiple R-squared:  0.7702, Adjusted R-squared:  0.7242
## F-statistic: 16.76 on 7 and 35 DF,  p-value: 1.766e-09
```

```
summary(lmodf4)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (lbph >= -1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44079 -0.28482  0.03428  0.36222  1.70450
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.522974   1.682782   0.905   0.370
## lcavol      0.624071   0.143729   4.342 7.92e-05 ***
## lweight     0.301422   0.233881   1.289   0.204
## age        -0.009231   0.017632  -0.524   0.603
## lbph       -0.018406   0.154367  -0.119   0.906
## svi         0.466097   0.319606   1.458   0.152
## lcp        -0.150206   0.126628  -1.186   0.242
## gleason    -0.077713   0.205455  -0.378   0.707
## pgg45       0.007021   0.005870   1.196   0.238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7121 on 45 degrees of freedom
## Multiple R-squared:  0.5199, Adjusted R-squared:  0.4345
## F-statistic: 6.091 on 8 and 45 DF,  p-value: 2.62e-05
```

```
summary(lmodf5)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (lcp <= -1))
##
## Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -1.25980 -0.34171 -0.03454  0.52731  1.56863
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.108126   2.532628   0.438  0.6643
## lcavol       0.594068   0.103340   5.749 1.38e-06 ***
## lweight      0.550927   0.206062   2.674  0.0111 *
## age         -0.009077   0.014622  -0.621  0.5385
## lbph        0.095935   0.085901   1.117  0.2713
## svi         0.916434   0.736094   1.245  0.2210
## lcp          NA         NA         NA     NA
## gleason     -0.186394   0.372433  -0.500  0.6197
## pgg45       0.020223   0.014014   1.443  0.1574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6864 on 37 degrees of freedom
## Multiple R-squared:  0.6705, Adjusted R-squared:  0.6082
## F-statistic: 10.76 on 7 and 37 DF,  p-value: 2.615e-07
```

```
summary(lmodf6)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate, subset = (lcp >= -1))
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.67941 -0.34127  0.03587  0.43100  1.59686
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.991685   1.862780   1.606  0.11559
## lcavol       0.549445   0.200644   2.738  0.00895 **
## lweight      0.304117   0.323037   0.941  0.35174
## age         -0.034035   0.018494  -1.840  0.07263 .
## lbph        0.109918   0.086906   1.265  0.21276
## svi         0.874348   0.284323   3.075  0.00365 **
## lcp         -0.151808   0.174817  -0.868  0.39001
## gleason     -0.051571   0.201626  -0.256  0.79934
## pgg45       0.003581   0.004948   0.724  0.47320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7359 on 43 degrees of freedom
## Multiple R-squared:  0.5406, Adjusted R-squared:  0.4551
## F-statistic: 6.325 on 8 and 43 DF,  p-value: 2.089e-05
```