STAT 210
Applied Statistics and Data Analysis
Problem List 10 - Solution
(Due on Week 11)

Exercise 1

Using the `sat` dataset in the `faraway` package, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

```r
library(faraway)
library(car)
str(sat)
```

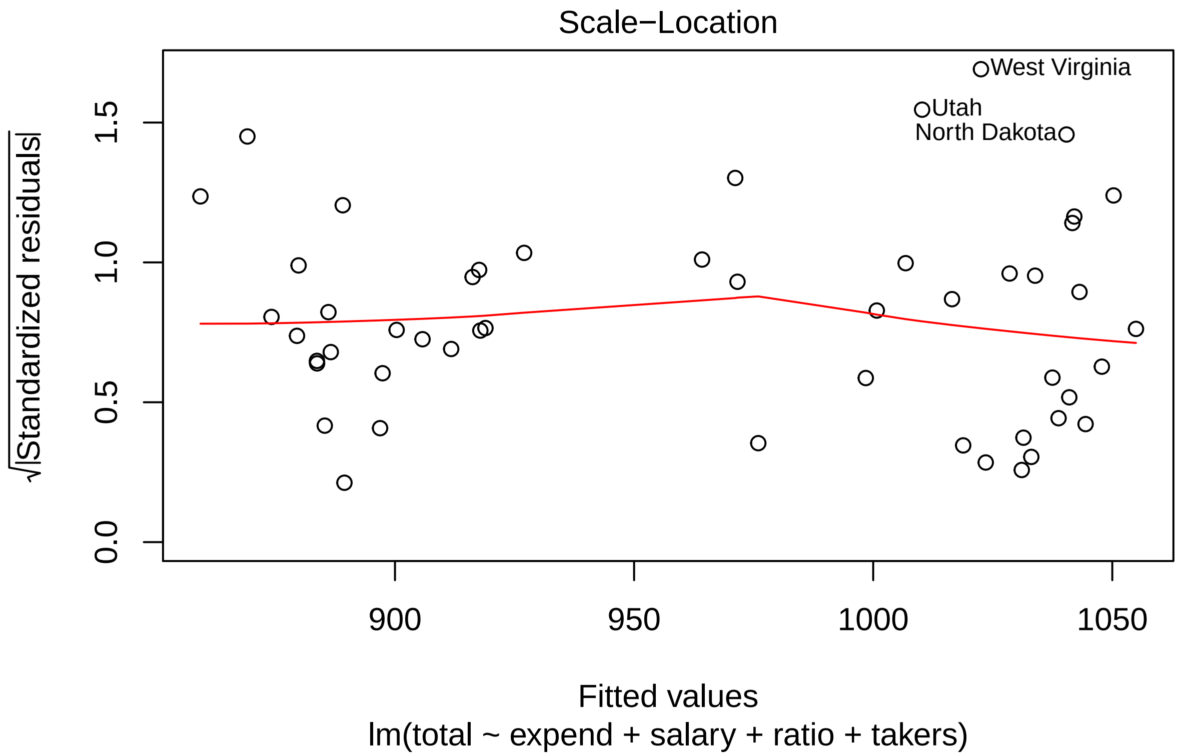
```
## 'data.frame':    50 obs. of  7 variables:
##  $ expend: num  4.41 8.96 4.78 4.46 4.99 ...
##  $ ratio : num  17.2 17.6 19.3 17.1 24 18.4 14.4 16.6 19.1 16.3 ...
##  $ salary: num  31.1 48 32.2 28.9 41.1 ...
##  $ takers: int  8 47 27 6 45 29 81 68 48 65 ...
##  $ verbal: int  491 445 448 482 417 462 431 429 420 406 ...
##  $ math  : int  538 489 496 523 485 518 477 468 469 448 ...
##  $ total : int  1029 934 944 1005 902 980 908 897 889 854 ...
```

```r
q4.mod <- lm(total ~ expend + salary + ratio + takers, data = sat)
summary(q4.mod)
```

```
##
## 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
```

(a) Check the constant variance assumption for the errors.

```r
plot(q4.mod, which = 3)
```



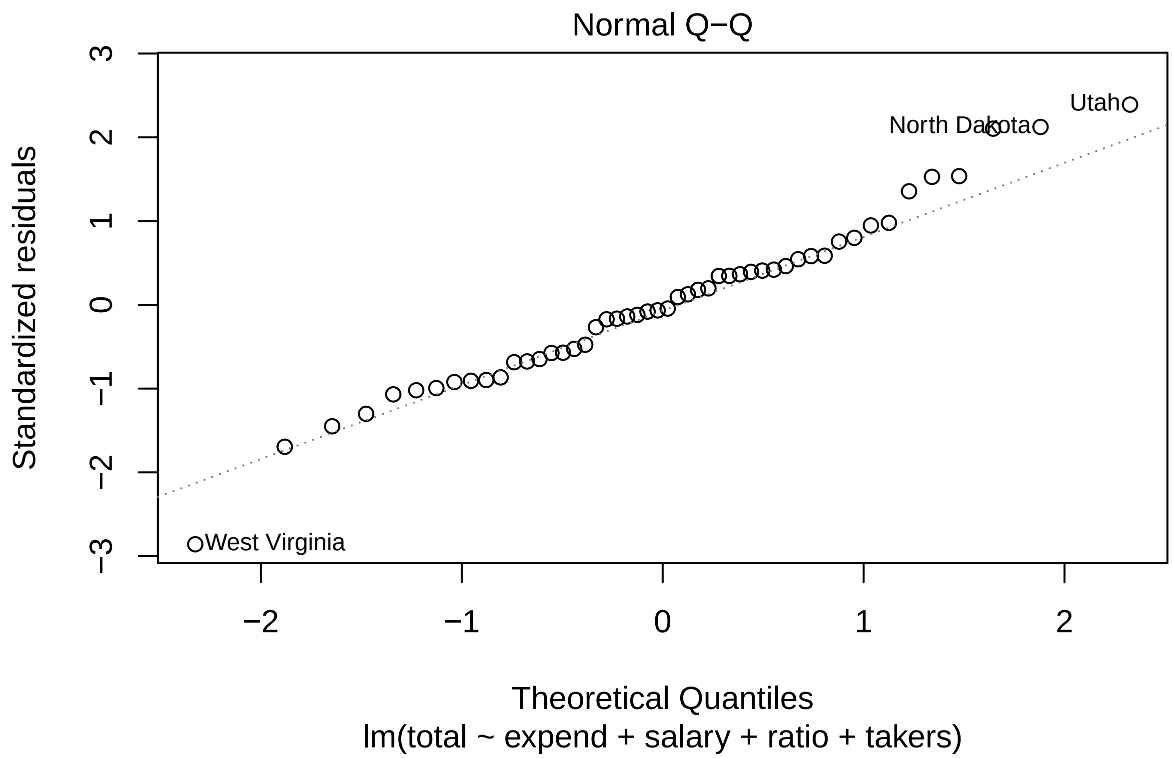

```r
ncvTest(q4.mod)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6972119, Df = 1, p = 0.40372
```

The assumption seems satisfied.

(b) Check the normality assumption.

```r
plot(q4.mod, which = 2)
```



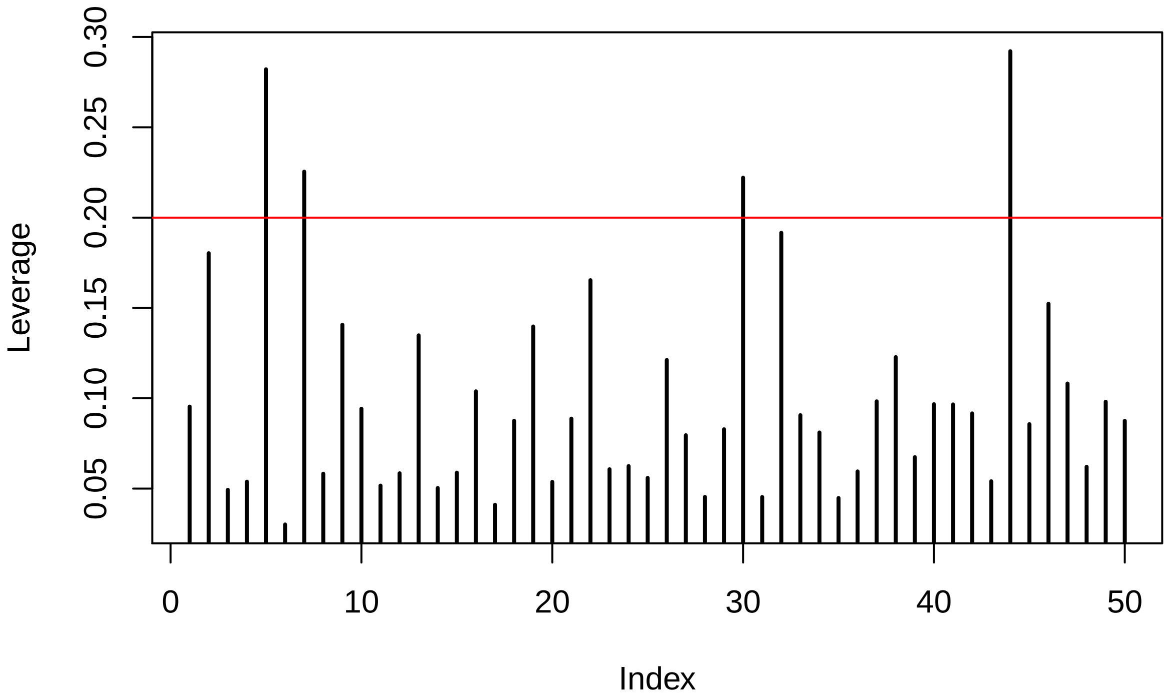

```r
shapiro.test(q4.mod$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  q4.mod$residuals
## W = 0.97691, p-value = 0.4304
```

The plot and test say that the normality assumption is satisfied.

(c) Check for large leverage points.

```r
plot(hatvalues(q4.mod), type = 'h', lwd=2, ylab='Leverage')
abline(h=0.2, col='red')
```



There are four points above the red line at $2p/n = 10/20 = 0.2$ that correspond to

```
high.lev <- (1:50)[hatvalues(q4.mod)>0.2]
dimnames(sat)[[1]][high.lev]
```

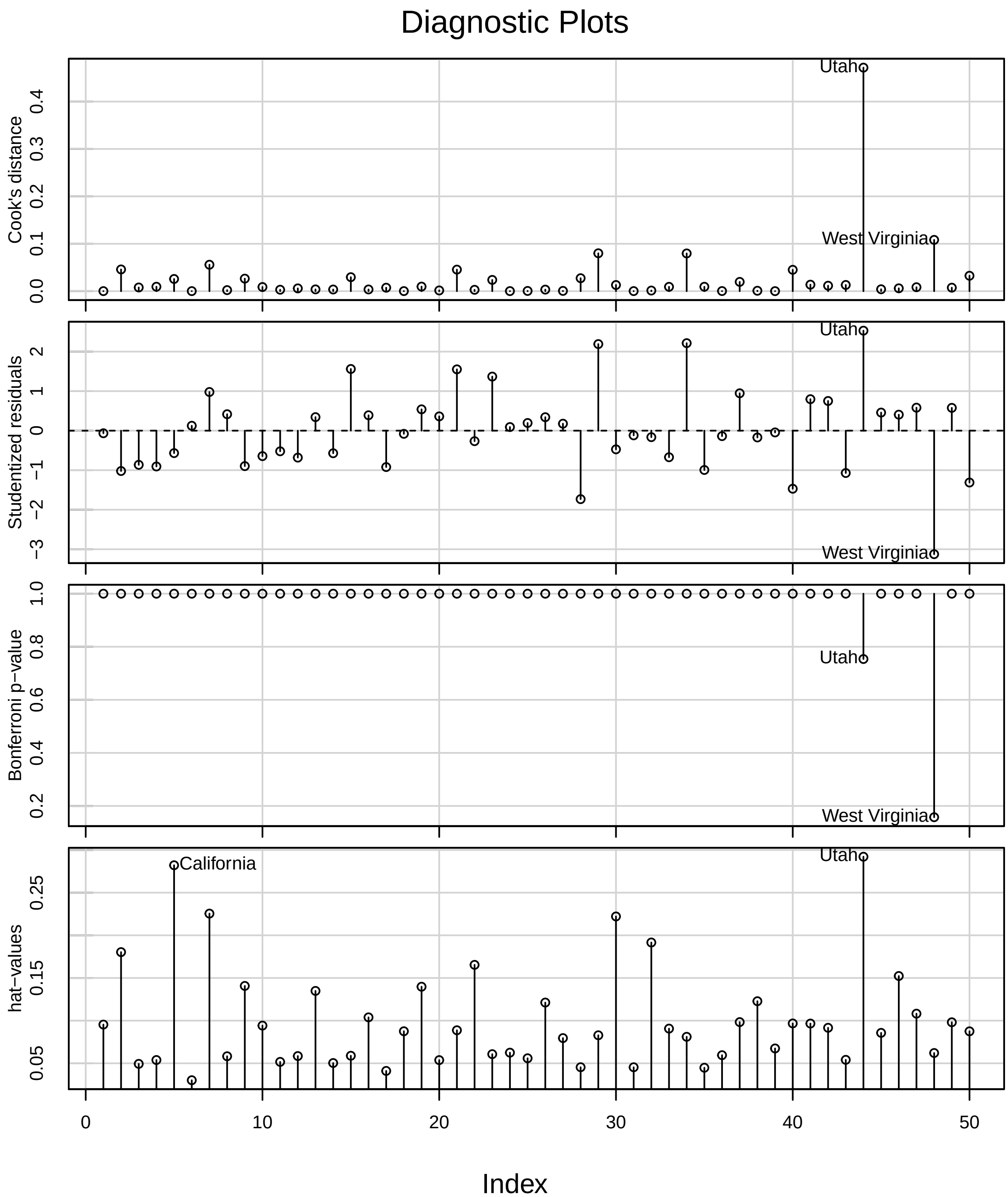
```
## [1] "California"  "Connecticut" "New Jersey"  "Utah"
```

The two highest values correspond to Utah and California.

(d) Check for outliers.

The following plots are from the `car` package. There are other alternatives.

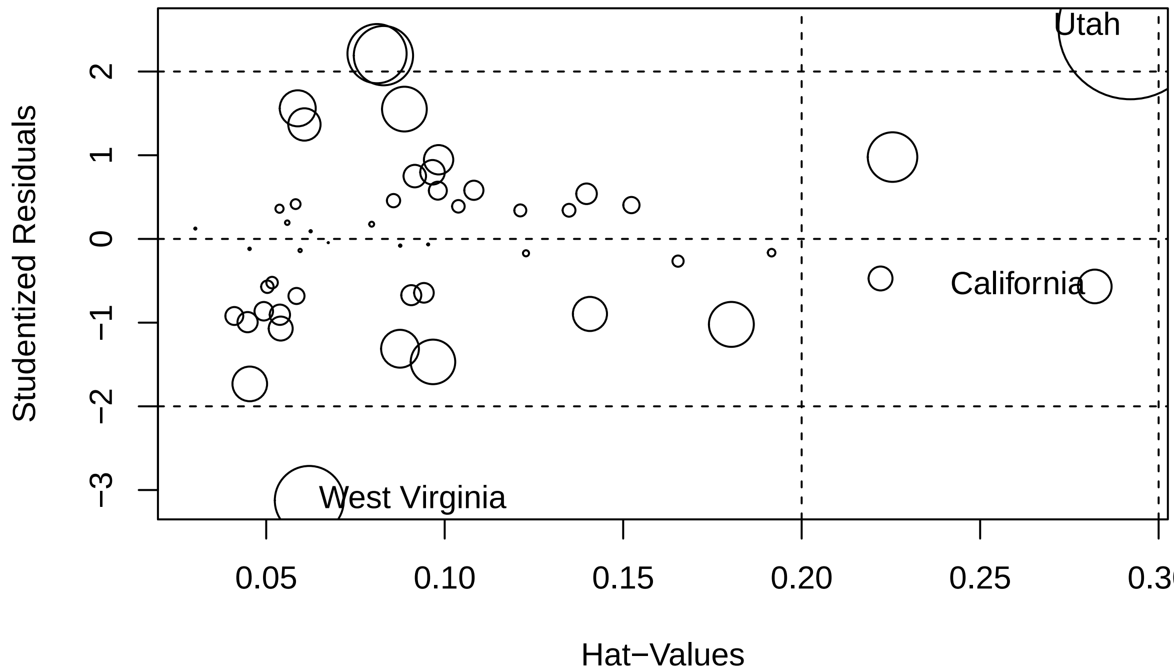
```
influenceIndexPlot(q4.mod)
```



In these plots West Virginia also appears as a possible outlier, even though the p-value in the Bonferroni graph is not small enough. However, this test is conservative and this may well be an outlier.

(e) Check for influential points.

```
influencePlot(q4.mod)
```




```
##                   StudRes        Hat      CookD
## California    -0.5676458 0.28211791 0.02571304
## Utah           2.5295873 0.29211280 0.47152866
## West Virginia -3.1244283 0.06206536 0.10813954
```

These are the three points that should be checked. As an example (but this was not required) we can look at the changes in the model produced by leaving out these points. These differences are included in the DFBETAS vector that can be extracted from the model using the function `dfbeta()`. For instance, for Utah the DFBETAS are

```
dfbeta(q4.mod)[44,]
```

```
##  (Intercept)       expend       salary        ratio       takers
## -47.87443743   5.40533354  -1.45851225   4.01491196   0.02632387
```

and the coefficients for the complete model are

```
coef(q4.mod)
```

```
## (Intercept)      expend      salary       ratio      takers
## 1045.971536    4.462594    1.637917   -3.624232   -2.904481
```

We see that the changes in some of the coefficients are quite significant.

For West Virginia

```
dfbeta(q4.mod)[48,]
```

```
##  (Intercept)       expend       salary        ratio       takers
## -11.70596970  -2.89670620   0.55038543   0.31550287   0.06791298
```

and for California

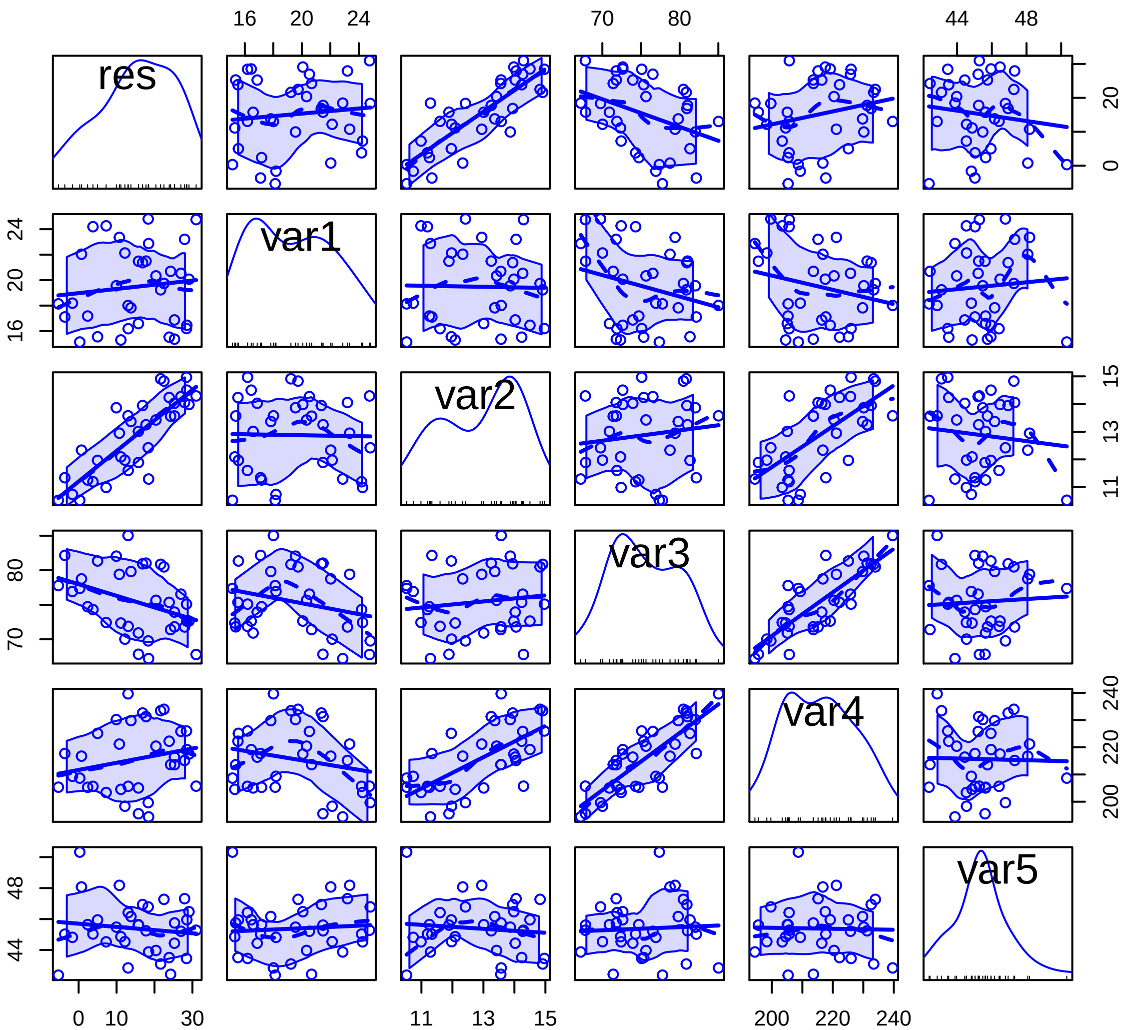
```
dfbeta(q4.mod)[5,]
```

```
## (Intercept)      expend      salary       ratio      takers
##  9.31992536  1.26239336 -0.30738250 -0.36782899 -0.00878391
```

Exercise 2

The data set `data_q2` has six variables. Find a minimal adequate model for `res` in terms of the other variables (without interactions). In your answer include exploratory analysis, variable selection and residual analysis. In each step **justify clearly the reason for your decision**. Give a prediction of `res` using your model for a subject with values (`var1`,`var2`,`var3`,`var4`,`var5`) $= (16.1, 14.0, 66.8, 202, 45.4)$, including confidence and prediction intervals.

```
q2.df <- read.table('data_q2.txt')
scatterplotMatrix(q2.df)
```



Variables 3 and 4 appear to have a high correlation. Fit a complete model

```
mod1 <- lm(res ~ ., data = q2.df)
summary(mod1)
```

```
##
## Call:
## lm(formula = res ~ ., data = q2.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9319 -1.3800 -0.0179  1.3305  2.9911
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.571929  11.823469  -0.133   0.8952
## var1         0.089258   0.120030   0.744   0.4631
## var2         2.024320   0.844324   2.398   0.0232 *
## var3        -2.899361   0.321081  -9.030 6.34e-10 ***
## var4         0.962949   0.163444   5.892 2.15e-06 ***
## var5        -0.001003   0.209053  -0.005   0.9962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.055 on 29 degrees of freedom
## Multiple R-squared:  0.9664, Adjusted R-squared:  0.9606
## F-statistic: 166.8 on 5 and 29 DF,  p-value: < 2.2e-16
```

Variables 2, 3 and 4 appear significant but we need to check for collinearity

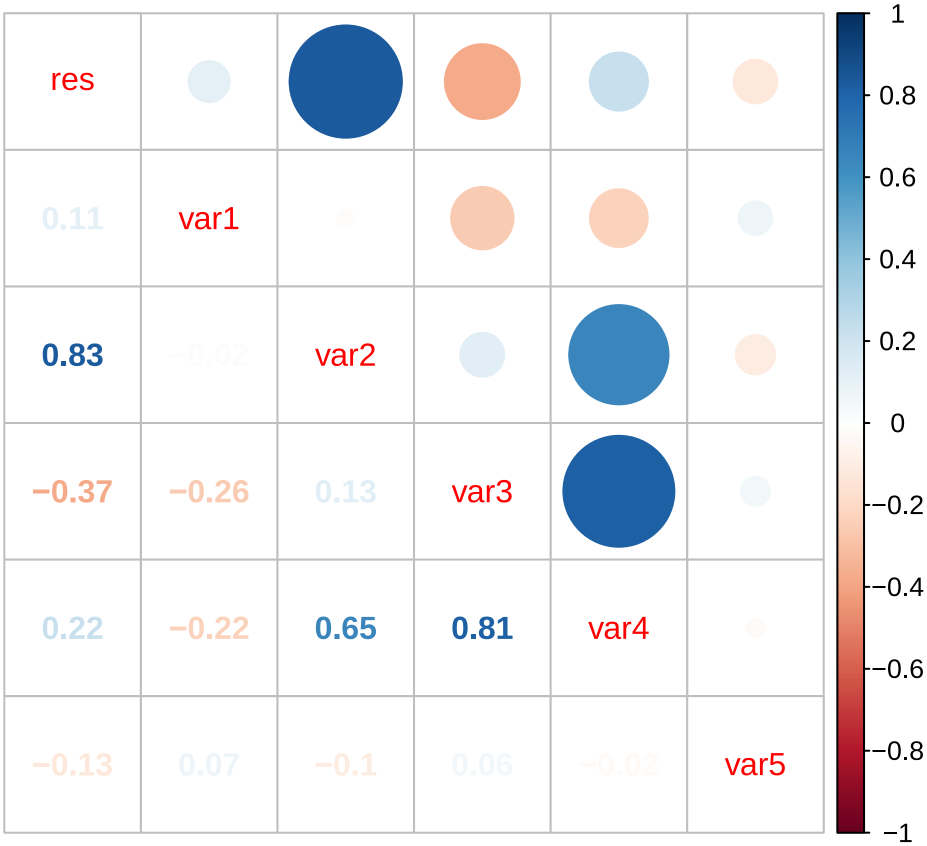
```r
vif(mod1)
```

```
##      var1      var2      var3      var4      var5
##  1.092813 10.789125 18.310165 31.548294  1.026067
```

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
cor.q2 <- cor(q2.df)
corrplot.mixed(cor.q2)
```



We see the Variance Inflation Factors for variables 4, 3 and 2 are large and the correlation matrix also shows large values. Since vif is largest for variable 4, we try dropping it from the model

```
mod2 <- update(mod1, .~. - var4)
summary(mod2)
```

```
##
## Call:
## lm(formula = res ~ var1 + var2 + var3 + var5, data = q2.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1273 -1.4456 -0.4035  1.6708  5.9888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.01610   16.99018   0.590    0.560
## var1         0.01580    0.17397   0.091    0.928
## var2         6.75536    0.38017  17.769  < 2e-16 ***
## var3        -1.06528    0.11459  -9.296 2.43e-10 ***
## var5        -0.03968    0.30450  -0.130    0.897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.995 on 30 degrees of freedom
## Multiple R-squared:  0.9262, Adjusted R-squared:  0.9163
## F-statistic: 94.07 on 4 and 30 DF,  p-value: < 2.2e-16
```

```
vif(mod2)
```

```
##     var1     var2     var3     var5
## 1.081023 1.029974 1.098158 1.025054
```

Now variables 2 and 3 are significant and the vif have decreased to normal levels. `var1` has the largest *p*-value, so we drop it from the model

```
mod3  <- update(mod2, .~.-var1)
summary(mod3)
```

```
##
## Call:
## lm(formula = res ~ var2 + var3 + var5, data = q2.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0844 -1.4346 -0.3921  1.6689  6.0360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.40080   16.18856   0.642    0.525
## var2         6.75621    0.37392  18.068  < 2e-16 ***
## var3        -1.06802    0.10875  -9.821 4.92e-11 ***
## var5        -0.03706    0.29824  -0.124    0.902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.946 on 31 degrees of freedom
```

```
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.919
## F-statistic: 129.6 on 3 and 31 DF,  p-value: < 2.2e-16
```

Finally, we drop `var5`, which is also non-significant.

```r
mod4 <- update(mod3, .~.-var5)
summary(mod4)
```

```
##
## Call:
## lm(formula = res ~ var2 + var3, data = q2.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1779 -1.3647 -0.4672  1.7131  6.0448
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7246     8.8101    0.99    0.329
## var2          6.7614     0.3658   18.48  < 2e-16 ***
## var3         -1.0690     0.1068  -10.01 2.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.901 on 32 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9215
## F-statistic: 200.5 on 2 and 32 DF,  p-value: < 2.2e-16
```

We can compare the initial model without `var4` with the final model using an anova

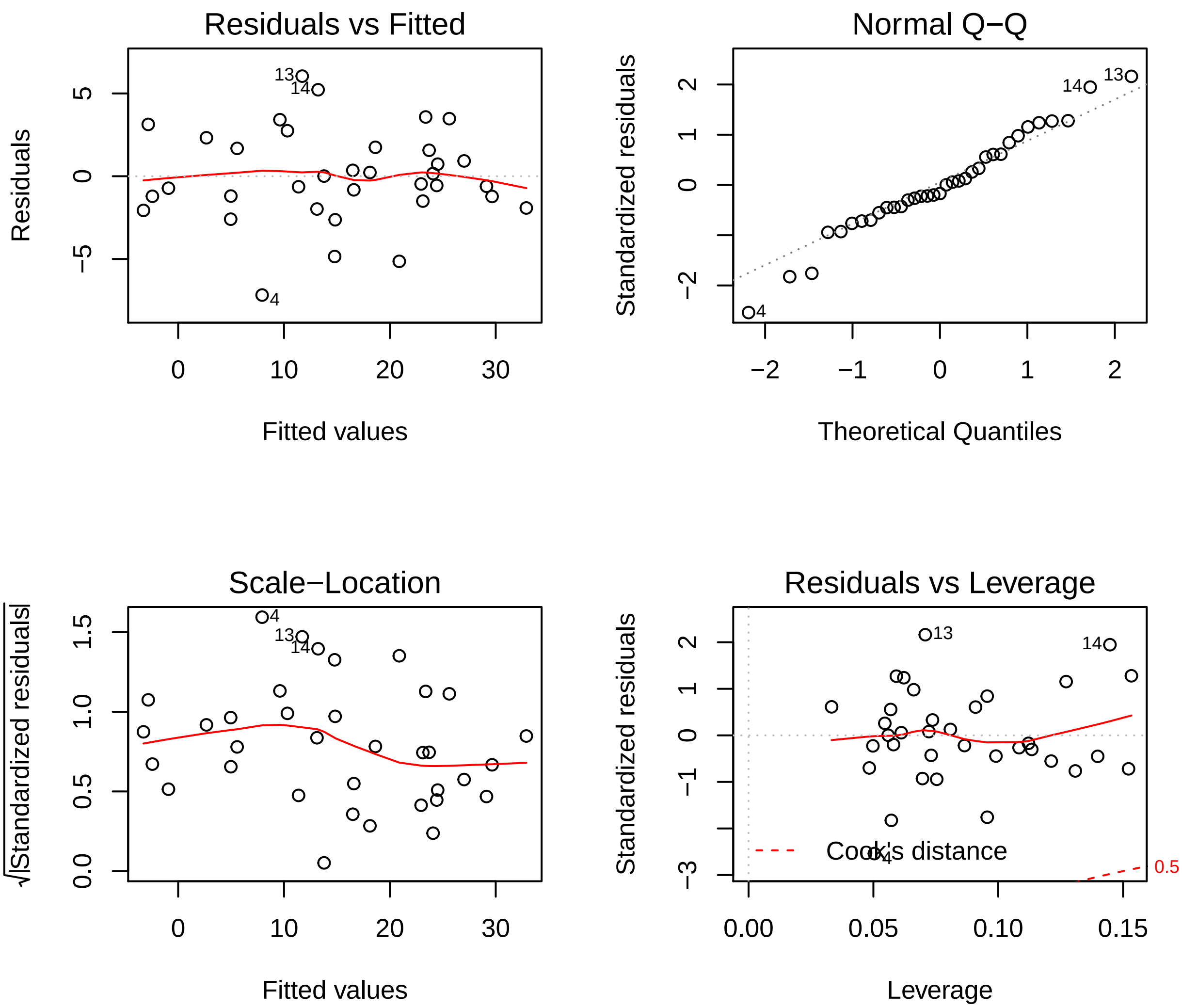
```r
anova(mod2,mod4)
```

```
## Analysis of Variance Table
##
## Model 1: res ~ var1 + var2 + var3 + var5
## Model 2: res ~ var2 + var3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 269.03
## 2     32 269.24 -2  -0.20802 0.0116 0.9885
```

and we keep the simpler model.

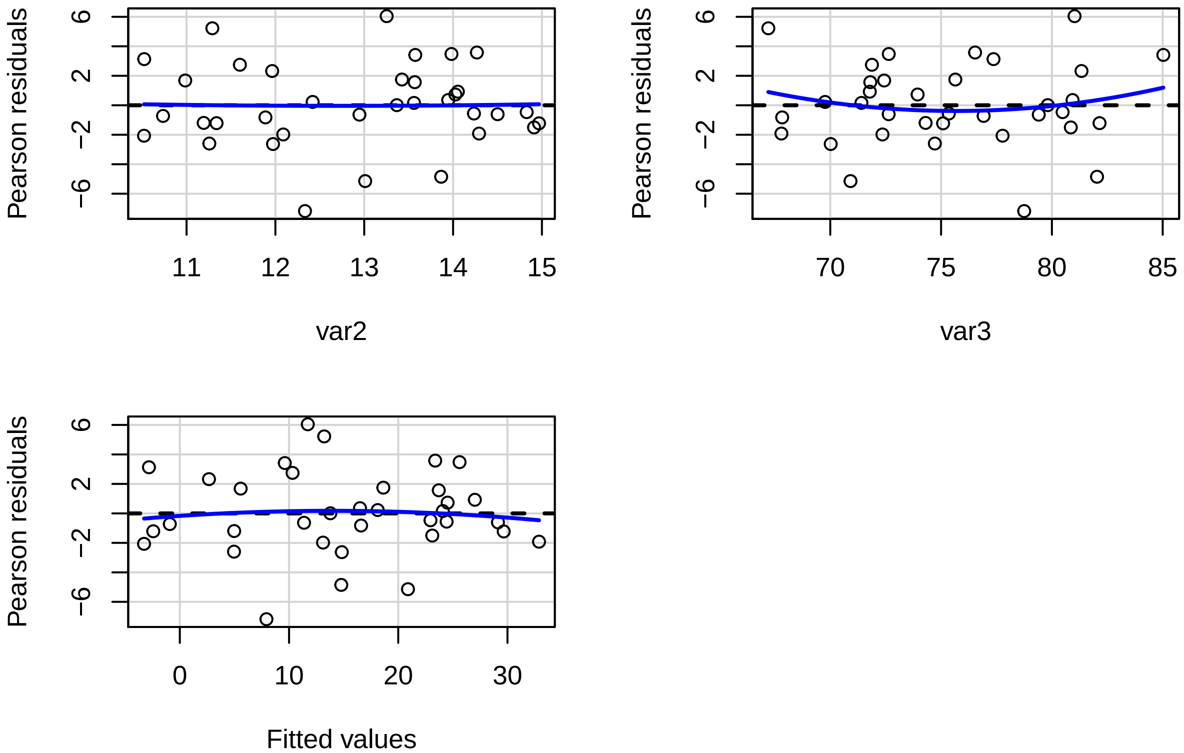
We look now at the residual plots

```r
par(mfrow=c(2,2))
plot(mod4)
```




```
par(mfrow=c(1,1))
```

```
residualPlots(mod4)
```




```
##            Test stat Pr(>|Test stat|)
## var2          0.0668           0.9472
## var3          0.7876           0.4369
## Tukey test   -0.3566           0.7214
```

These graphs show that the usual hypothesis are satisfied and the fit is good.

```r
ncvTest(mod4)
```

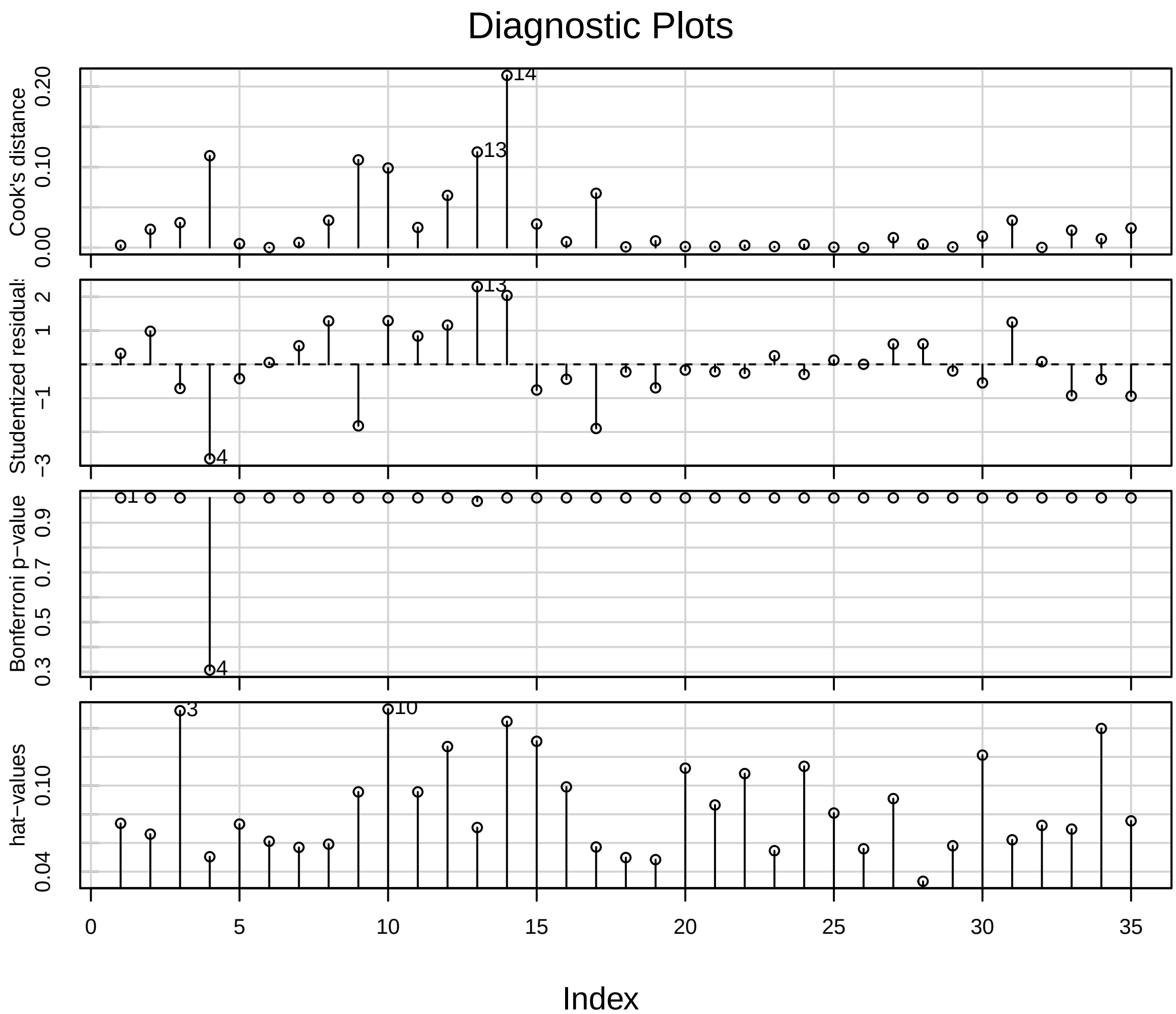
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8837723, Df = 1, p = 0.34717
```

```r
shapiro.test(rstandard(mod4))
```

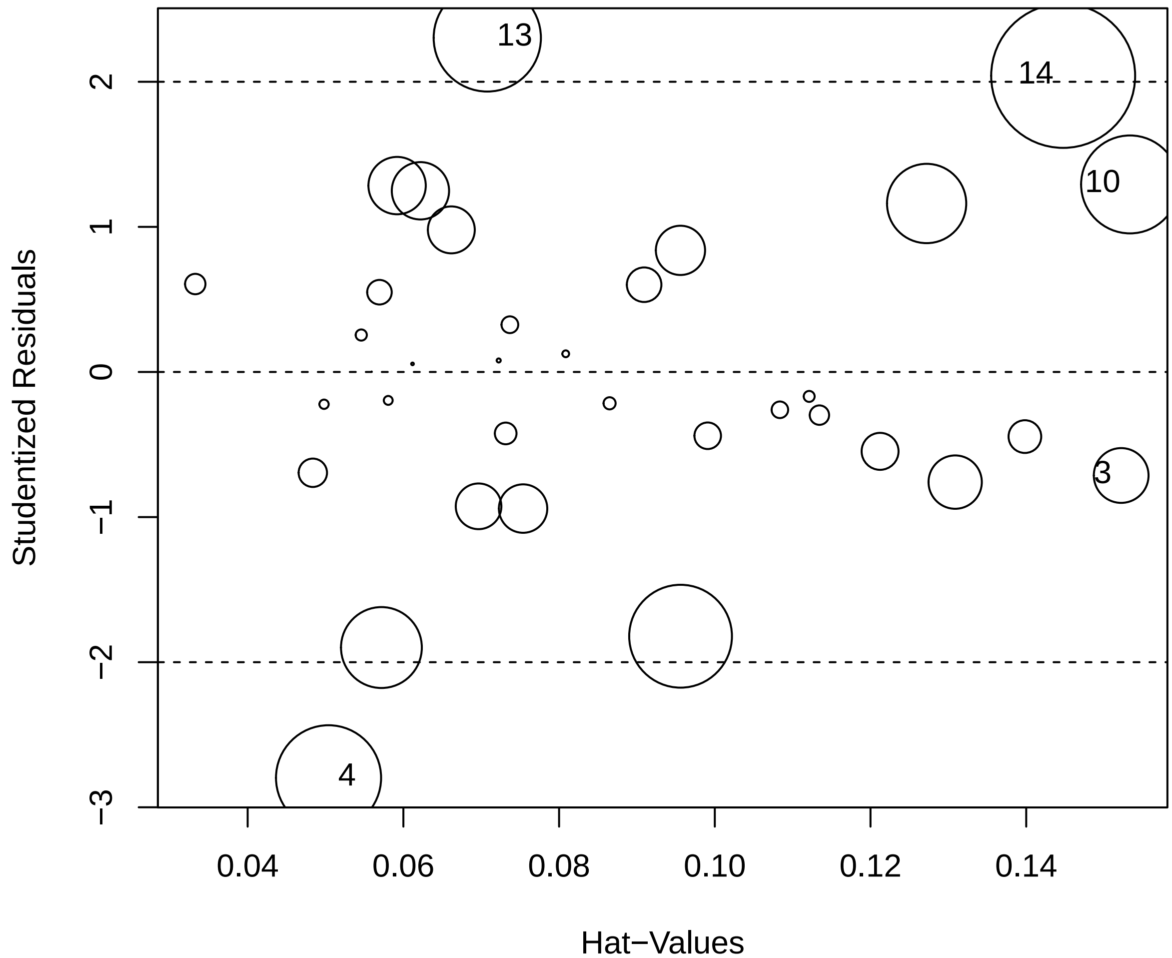
```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod4)
## W = 0.97876, p-value = 0.7184
```

The tests confirm this.

```r
influenceIndexPlot(mod4)
```




```
influencePlot(mod4)
```




```
##      StudRes        Hat      CookD
## 3  -0.713237 0.15219727 0.03091562
## 4  -2.797092 0.05038808 0.11405823
## 10  1.292615 0.15333481 0.09879518
## 13  2.302600 0.07077735 0.11866159
## 14  2.041033 0.14474618 0.21385541
```

In the influence plots we see that none of the points has high leverage or large Cook's distance. A few values for the standardized residuals are large (above 2 in absolute value). The worse point seems to be 14. We check the effect of this point using DFBETAS

```r
dfbeta(mod4)[14,]
```

```
## (Intercept)        var2        var3
##  6.47470164 -0.12433035 -0.06230186
```

and the coefficients for the complete model are

```r
coef(mod4)
```

```
## (Intercept)        var2        var3
##    8.724599    6.761377   -1.068982
```

We see that the changes are not important. For the other two points:

```r
dfbeta(mod4)[13,]
```

```
## (Intercept)        var2        var3
## -3.64928316  0.01725356  0.04790045
```

```r
dfbeta(mod4)[10,]
```

```
## (Intercept)        var2        var3
## -3.99813400  0.02176309  0.05081904
```

To give a prediction we only need the values for `var2 = 14.0` and `var3 = 66.8`.

```r
new.data <- data.frame(var2=14.0, var3=66.8)
predict(mod4,newdata = new.data, interval = c('confidence'))
```

```
##       fit      lwr      upr
## 1 31.9759 29.60508 34.34672
```

```r
predict(mod4,newdata = new.data, interval = c('prediction'))
```

```
##       fit      lwr      upr
## 1 31.9759 25.60957 38.34223
```

Exercise 3

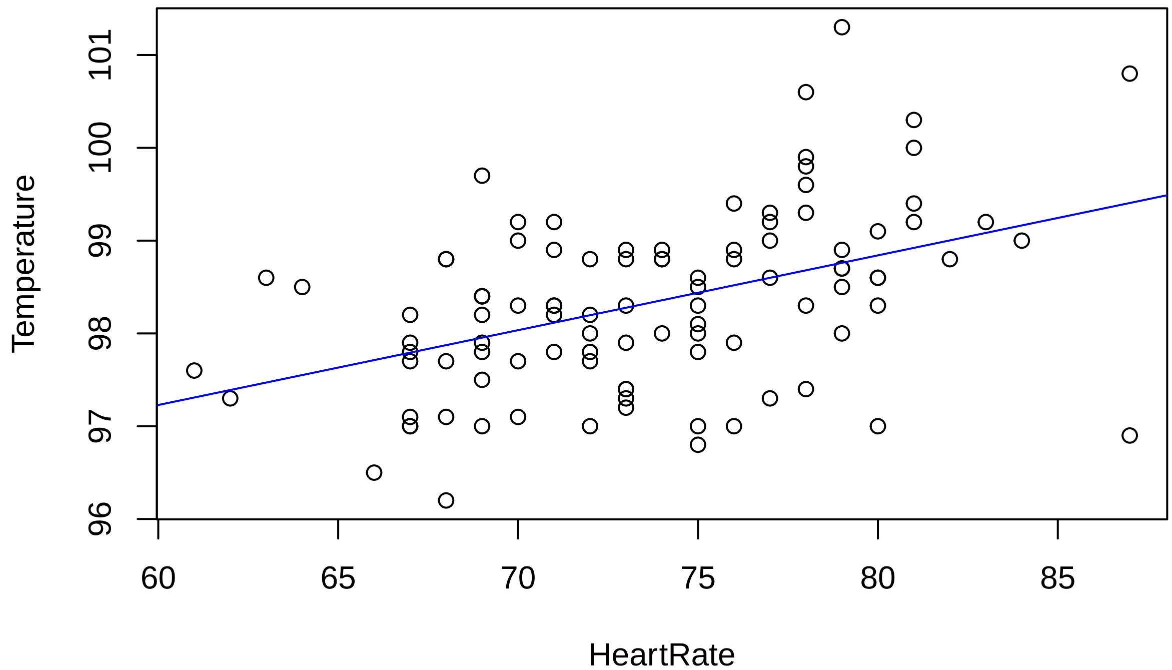
We want to examine the relationship between body temperature Y and heart rate X. Further, we would like to use heart rate to predict the body temperature.

(a) Use the `BodyTemperature.txt` data set to build a simple linear regression model for body temperature using heart rate as the predictor.

```r
bodytemp <- read.table('BodyTemperature.txt', header = TRUE)
str(bodytemp)
```

```
## 'data.frame':    100 obs. of  4 variables:
##  $ Gender     : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 1 1 1 2 ...
##  $ Age        : int  33 32 42 33 26 37 32 45 31 49 ...
##  $ HeartRate  : int  69 72 68 75 68 79 71 73 77 81 ...
##  $ Temperature: num  97 98.8 96.2 97.8 98.8 ...
```

```r
bodyt.mod <- lm(Temperature ~ HeartRate, data = bodytemp)
plot(Temperature ~ HeartRate, data = bodytemp)
abline(bodyt.mod, col = 'blue')
```




```r
summary(bodyt.mod)
```

```
##
## Call:
## lm(formula = Temperature ~ HeartRate, data = bodytemp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50562 -0.46473  0.00543  0.48943  2.53943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 92.39068    1.20144  76.900  < 2e-16 ***
## HeartRate    0.08063    0.01627   4.956 3.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.86 on 98 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.1923
## F-statistic: 24.56 on 1 and 98 DF,  p-value: 3.011e-06
```

(b) Interpret the estimate of regression coefficient and examine its statistical significance.

The coefficient is 0.08063 and is statitically significant at the usual levels. It means that an increase of 1 unit in heart rate produces an increase of 0.08063 degrees in body temperature.

(c) Find the 95% confidence interval for the regression coefficient.

```r
confint(bodyt.mod)
```

```
##                   2.5 %     97.5 %
## (Intercept) 90.00646174 94.7748998
## HeartRate    0.04834668  0.1129164
```

(d) Find the value of R^2 and show that it is equal to sample correlation coefficient.

```r
summary(bodyt.mod)$r.squared
```

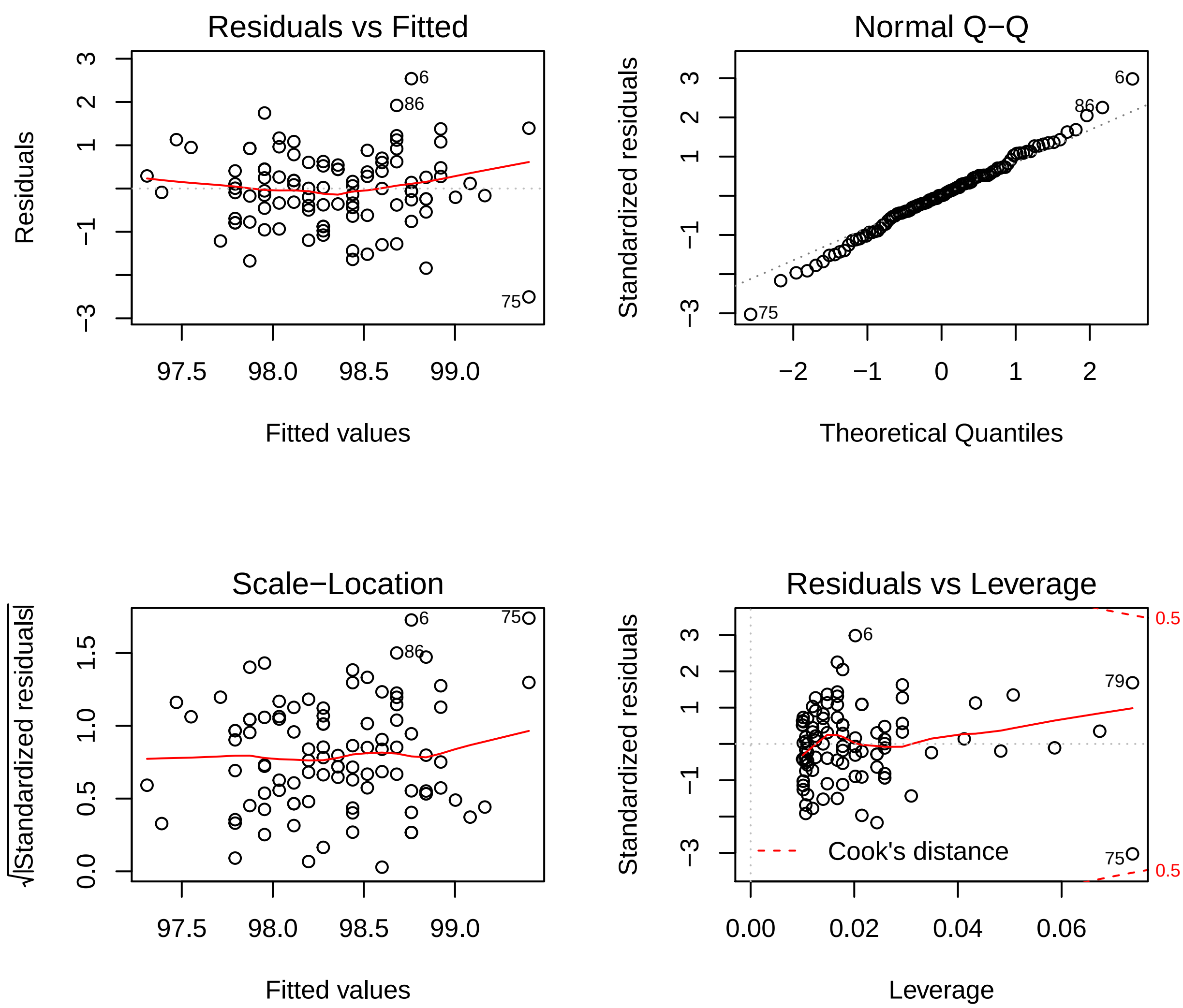
```
## [1] 0.2004181
```

```r
with(bodytemp ,cor(Temperature,HeartRate)^2)
```

```
## [1] 0.2004181
```

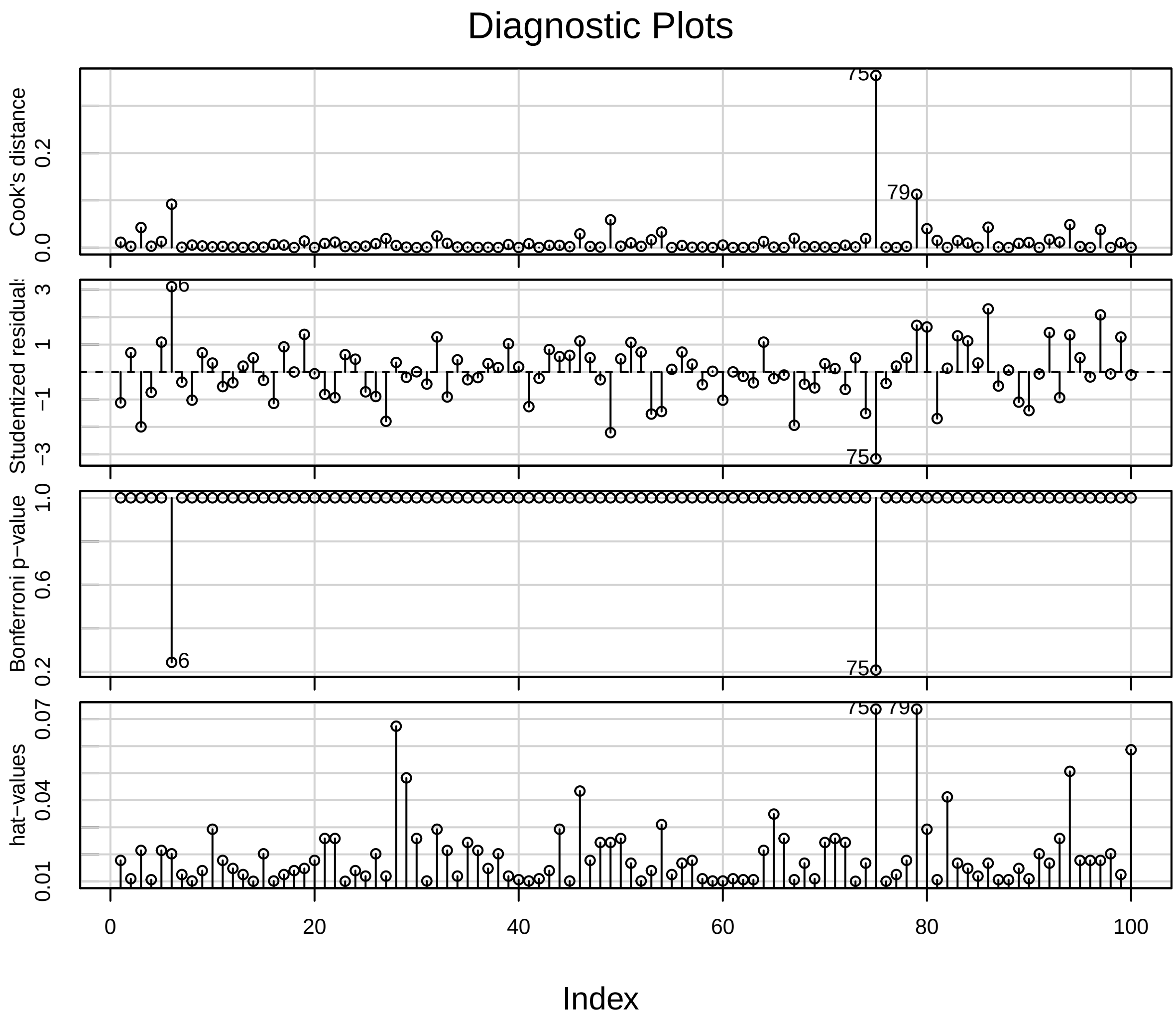
(e) Create simple diagnostic plots for your model and identify possible outliers.

```r
par(mfrow=c(2,2))
plot(bodyt.mod)
```

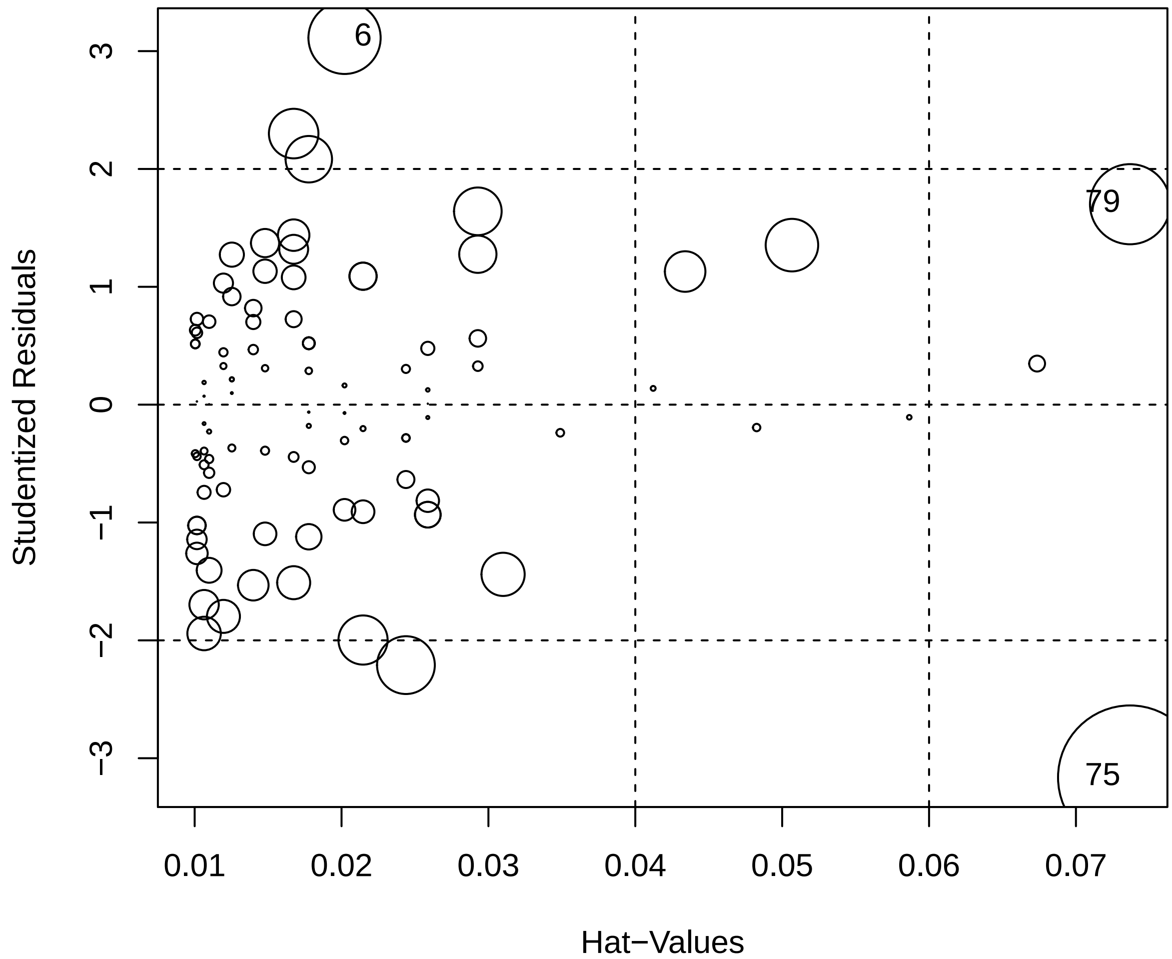


Points 6, and 75 are flagged in all diagnostic graphs.

```r
influenceIndexPlot(bodyt.mod)
```




```r
influencePlot(bodyt.mod)
```




```
##       StudRes        Hat      CookD
## 6   3.112503 0.02020441 0.09175127
## 75 -3.163181 0.07368203 0.36445066
## 79  1.700792 0.07368203 0.11286675
```

Look at the effect of point 79.

```
dfbeta(bodyt.mod)[79,]
```

```
##  (Intercept)    HeartRate
## -0.514259842  0.007185891
```

```
coef(bodyt.mod)
```

```
## (Intercept)   HeartRate
## 92.39068078  0.08063154
```

```
dfbeta(bodyt.mod)[75,]
```

```
## (Intercept)   HeartRate
##   0.9241000  -0.0129127
```

```
dfbeta(bodyt.mod)[6,]
```

```
##  (Intercept)    HeartRate
## -0.338901978  0.004952755
```

(f) If someone's heart rate is 75, what would be your estimate of this person's body temperature?

Using `predict` and adding a confidence interval

```
predict(bodyt.mod,list(HeartRate = 75), interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 98.43805 98.26198 98.61411
```

Using the coefficients and the model formula:

```
coef(bodyt.mod)[1] + 75*coef(bodyt.mod)[2]
```

```
## (Intercept)
##    98.43805
```

We believe that gender might also be related to body temperature and could help us to predict its unknown values. For the tests in this section use $\alpha = 0.1$.

(g) Use the "BodyTemperature.txt" data set to build a multiple linear regression model for body temperature using heart rate and gender as predictors.

```
bodyt.mod2 <- lm(Temperature ~ HeartRate*Gender, data = bodytemp)
anova(bodyt.mod2)
```

```
## Analysis of Variance Table
##
## Response: Temperature
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## HeartRate         1 18.168 18.1679 24.8421 2.754e-06 ***
## Gender            1  2.251  2.2505  3.0773   0.08258 .
## HeartRate:Gender  1  0.024  0.0235  0.0321   0.85810
## Residuals        96 70.208  0.7313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(bodyt.mod2)
```

```
##
## Call:
## lm(formula = Temperature ~ HeartRate * Gender, data = bodytemp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33872 -0.47940 -0.00335  0.53645  2.69775
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       92.183303   1.854876  49.698  < 2e-16 ***
## HeartRate          0.085457   0.025214   3.389  0.00102 **
## GenderM            0.133777   2.428067   0.055  0.95618
## HeartRate:GenderM -0.005898   0.032899  -0.179  0.85810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8552 on 96 degrees of freedom
## Multiple R-squared:  0.2255, Adjusted R-squared:  0.2013
## F-statistic: 9.317 on 3 and 96 DF,  p-value: 1.822e-05
```

The interaction is not significant so we fit a model without interaction.

```r
bodyt.mod3 <- lm(Temperature ~ HeartRate + Gender, data = bodytemp)
anova(bodyt.mod3)
```

```
## Analysis of Variance Table
##
## Response: Temperature
##           Df Sum Sq Mean Sq F value    Pr(>F)
## HeartRate  1 18.168 18.1679 25.0925 2.452e-06 ***
## Gender     1  2.251  2.2505  3.1083   0.08104 .
## Residuals 97 70.232  0.7240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(bodyt.mod3)
```

```
##
## Call:
## lm(formula = Temperature ~ HeartRate + Gender, data = bodytemp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37056 -0.48862 -0.00963  0.53575  2.68538
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 92.43764    1.18902  77.743  < 2e-16 ***
## HeartRate    0.08199    0.01612   5.088 1.77e-06 ***
## GenderM     -0.30044    0.17041  -1.763    0.081 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8509 on 97 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2093
## F-statistic:  14.1 on 2 and 97 DF,  p-value: 4.212e-06
```

In this model `Gender` is marginally significant. At the $\alpha = 0.05$ level we would keep the simple regression model using only `HeartRate` and then the rest of the question will have the same answer as in the first part. However, if we choose $\alpha = 0.1$ then `Gender` is significant and the answers are different.

We answer the next four questions using the additive model with `HeartRate` and `Gender` as variables.

(h) How much R^2 did increase compared the above simple linear regression model? The first value corresponds to the additive model, the second to the simple linear model.

```r
summary(bodyt.mod3)$r.squared;summary(bodyt.mod)$r.squared
```

```
## [1] 0.2252448
```

```
## [1] 0.2004181
```

(i) Explain the estimates of regression coefficients in plain language.

The additive model has a common slope of 0.08199 for both genders but the intercepts are different. For females the intercept is 92.43764 while for males it is $92.43764 - 0.30044 = 92.1372$

(j) Find the 95% confidence intervals for regression coefficients.

```r
confint(bodyt.mod3)
```

```
##                   2.5 %      97.5 %
## (Intercept) 90.07776129 94.79751155
## HeartRate    0.05000874  0.11397659
## GenderM     -0.63865874  0.03777622
```

(k) If a woman's heart rate is 75, what would be your estimate of her body temperature? What would be your estimate of body temperature for a man whose heart rate is 75?

```r
predict(bodyt.mod3,list(HeartRate = 75, Gender='F'),
        interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 98.58709 98.34521 98.82896
```

```r
predict(bodyt.mod3,list(HeartRate = 75, Gender='M'),
        interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 98.28665 98.04292 98.53037
```

Exercise 4

For this question use the data set `uscrime` in the package `HH`. After loading the library, you need to run `data("uscrime")`. Do not mistake with `UScrime`. For this exercise, values for the variance inflation factor (vif) below 5 are considered acceptable. The following commands load the data:

```r
library(HH)
data("uscrime")
```

(a) Fit a multiple regression model for `R` using all the other variables except `State`. Look at the summary and variance inflation factors and comment.

We start by looking at the structure of the data set and plotting a scatterplot matrix.

```r
str(uscrime)
```

```
## 'data.frame':    47 obs. of  15 variables:
##  $ R    : num  79.1 163.5 57.8 196.9 123.4 ...
##  $ Age  : int  151 143 142 136 141 121 127 131 157 140 ...
##  $ S    : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed   : int  91 113 89 121 121 110 111 109 90 118 ...
##  $ Ex0  : int  58 103 45 149 109 118 82 115 65 71 ...
##  $ Ex1  : int  56 95 44 141 101 115 79 109 62 68 ...
##  $ LF   : int  510 583 533 577 591 547 519 542 553 632 ...
##  $ M    : int  950 1012 969 994 985 964 982 969 955 1029 ...
##  $ N    : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW   : int  301 102 219 80 30 44 139 179 286 15 ...
##  $ U1   : int  108 96 94 102 91 84 97 79 81 100 ...
##  $ U2   : int  41 36 33 39 20 29 38 35 28 24 ...
##  $ W    : int  394 557 318 673 578 689 620 472 421 526 ...
##  $ X    : int  261 194 250 167 174 126 168 206 239 174 ...
##  $ State: Factor w/ 47 levels "Alabama","Arizona",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```r
plot(uscrime[,-15])
```

We fit a linear regression model and look at the summary table and vif values.

```
lm1 <- lm(R ~ . - State, data = uscrime)
summary(lm1)
```

```
##
## Call:
## lm(formula = R ~ . - State, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.884 -11.923  -1.135  13.495  50.560
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.918e+02  1.559e+02  -4.438 9.56e-05 ***
## Age          1.040e+00  4.227e-01   2.460  0.01931 *
## S           -8.308e+00  1.491e+01  -0.557  0.58117
## Ed           1.802e+00  6.496e-01   2.773  0.00906 **
## Ex0          1.608e+00  1.059e+00   1.519  0.13836
## Ex1         -6.673e-01  1.149e+00  -0.581  0.56529
## LF          -4.103e-02  1.535e-01  -0.267  0.79087
## M            1.648e-01  2.099e-01   0.785  0.43806
## N           -4.128e-02  1.295e-01  -0.319  0.75196
## NW           7.175e-03  6.387e-02   0.112  0.91124
## U1          -6.017e-01  4.372e-01  -1.376  0.17798
## U2           1.792e+00  8.561e-01   2.093  0.04407 *
## W            1.374e-01  1.058e-01   1.298  0.20332
## X            7.929e-01  2.351e-01   3.373  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.94 on 33 degrees of freedom
## Multiple R-squared:  0.7692, Adjusted R-squared:  0.6783
## F-statistic: 8.462 on 13 and 33 DF,  p-value: 3.686e-07
```

```r
vif(lm1)
```

```
##       Age         S        Ed       Ex0       Ex1        LF         M
##  2.698021  4.876751  5.049442 94.633118 98.637233  3.677557  3.658444
##         N        NW        U1        U2         W         X
##  2.324326  4.123274  5.938264  4.997617  9.968958  8.409449
```

We see that many regressors are not significant and that some have very high vif values, particularly `Ex0` and `Ex1`.

(b) Use the function `stepAIC` in package `MASS` to get a reduced model. Get information about this model using `summary` and look at the variance inflation factors. **Comment on these results**.

```r
modelAIC <- stepAIC(lm1)
```

```
## Start:  AIC=301.66
## R ~ (Age + S + Ed + Ex0 + Ex1 + LF + M + N + NW + U1 + U2 + W +
##     X + State) - State
##
##        Df Sum of Sq   RSS    AIC
## - NW    1       6.1 15885 299.68
## - LF    1      34.4 15913 299.76
## - N     1      48.9 15928 299.81
## - S     1     149.4 16028 300.10
```

```
## - Ex1    1      162.3 16041 300.14
## - M      1      296.5 16175 300.53
## <none>                15879 301.66
## - W      1      810.6 16689 302.00
## - U1     1      911.5 16790 302.29
## - Ex0    1     1109.8 16988 302.84
## - U2     1     2108.8 17988 305.52
## - Age    1     2911.6 18790 307.57
## - Ed     1     3700.5 19579 309.51
## - X      1     5474.2 21353 313.58
##
## Step:  AIC=299.68
## R ~ Age + S + Ed + Ex0 + Ex1 + LF + M + N + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - LF    1      28.7 15913 297.76
## - N     1      48.6 15933 297.82
## - Ex1   1     156.3 16041 298.14
## - S     1     158.0 16043 298.14
## - M     1     294.1 16179 298.54
## <none>               15885 299.68
## - W     1     820.2 16705 300.05
## - U1    1     913.1 16798 300.31
## - Ex0   1    1104.3 16989 300.84
## - U2    1    2107.1 17992 303.53
## - Age   1    3365.8 19250 306.71
## - Ed    1    3757.1 19642 307.66
## - X     1    5503.6 21388 311.66
##
## Step:  AIC=297.76
## R ~ Age + S + Ed + Ex0 + Ex1 + M + N + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - N     1      62.2 15976 295.95
## - S     1     129.4 16043 296.14
## - Ex1   1     134.8 16048 296.16
## - M     1     276.8 16190 296.57
## <none>               15913 297.76
## - W     1     801.9 16715 298.07
## - U1    1     941.8 16855 298.47
## - Ex0   1    1075.9 16989 298.84
## - U2    1    2088.5 18002 301.56
## - Age   1    3407.9 19321 304.88
## - Ed    1    3895.3 19809 306.06
## - X     1    5621.3 21535 309.98
##
## Step:  AIC=295.95
## R ~ Age + S + Ed + Ex0 + Ex1 + M + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - S     1     104.4 16080 294.25
## - Ex1   1     123.3 16099 294.31
## - M     1     533.8 16509 295.49
## <none>               15976 295.95
```

```
## - W      1     748.7 16724 296.10
## - U1     1     997.7 16973 296.80
## - Ex0    1    1021.3 16997 296.86
## - U2     1    2082.3 18058 299.71
## - Age    1    3425.9 19402 303.08
## - Ed     1    3887.6 19863 304.19
## - X      1    5896.9 21873 308.71
##
## Step:  AIC=294.25
## R ~ Age + Ed + Ex0 + Ex1 + M + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - Ex1   1     171.5 16252 292.75
## - M     1     563.4 16643 293.87
## <none>              16080 294.25
## - W     1     734.7 16815 294.35
## - U1    1     906.0 16986 294.83
## - Ex0   1    1162.0 17242 295.53
## - U2    1    1978.0 18058 297.71
## - Age   1    3354.5 19434 301.16
## - Ed    1    4139.1 20219 303.02
## - X     1    6094.8 22175 307.36
##
## Step:  AIC=292.75
## R ~ Age + Ed + Ex0 + M + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - M     1     691.0 16942 292.71
## <none>              16252 292.75
## - W     1     759.0 17010 292.90
## - U1    1     921.8 17173 293.35
## - U2    1    2018.1 18270 296.25
## - Age   1    3323.1 19574 299.50
## - Ed    1    4005.1 20256 301.11
## - X     1    6402.7 22654 306.36
## - Ex0   1   11818.8 28070 316.44
##
## Step:  AIC=292.71
## R ~ Age + Ed + Ex0 + U1 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
## - U1    1     408.6 17351 291.83
## <none>              16942 292.71
## - W     1    1016.9 17959 293.45
## - U2    1    1548.6 18491 294.82
## - Age   1    4511.6 21454 301.81
## - Ed    1    6430.6 23373 305.83
## - X     1    8147.7 25090 309.16
## - Ex0   1   12019.6 28962 315.91
##
## Step:  AIC=291.83
## R ~ Age + Ed + Ex0 + U2 + W + X
##
##        Df Sum of Sq   RSS    AIC
```

```
## <none>              17351 291.83
## - W     1    1252.6 18604 293.11
## - U2    1    1628.7 18980 294.05
## - Age   1    4461.0 21812 300.58
## - Ed    1    6214.7 23566 304.22
## - X     1    8932.3 26283 309.35
## - Ex0   1   15596.5 32948 319.97
```

```r
summary(modelAIC)
```

```
##
## Call:
## lm(formula = R ~ Age + Ed + Ex0 + U2 + W + X, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.306 -10.209  -1.313   9.919  54.544
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -618.5028   108.2456  -5.714 1.19e-06 ***
## Age            1.1252     0.3509   3.207 0.002640 **
## Ed             1.8179     0.4803   3.785 0.000505 ***
## Ex0            1.0507     0.1752   5.996 4.78e-07 ***
## U2             0.8282     0.4274   1.938 0.059743 .
## W              0.1596     0.0939   1.699 0.097028 .
## X              0.8236     0.1815   4.538 5.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.83 on 40 degrees of freedom
## Multiple R-squared:  0.7478, Adjusted R-squared:   0.71
## F-statistic: 19.77 on 6 and 40 DF,  p-value: 1.441e-10
```

```r
vif(modelAIC)
```

```
##      Age       Ed      Ex0       U2        W        X
## 2.061942 3.061153 2.875709 1.381671 8.705602 5.559788
```

The `stepAIC` function has dropped many terms but we still have two vif values above 5, `W` and `X`.

(c) Starting with the model produced in (b), drop any variables that have a vif greater that 5 or non-significant p-value. Give a summary of your final model and write down the corresponding equation.

We drop `W`, that has the highest vif value

```r
modelAIC2 <- update(modelAIC, .~. - W)
vif(modelAIC2)
```

```
##      Age       Ed      Ex0       U2        X
## 1.997555 2.852816 1.796182 1.362551 3.479589
```

```r
summary(modelAIC2)
```

```
##
## Call:
## lm(formula = R ~ Age + Ed + Ex0 + U2 + X, data = uscrime)
##
## Residuals:
```

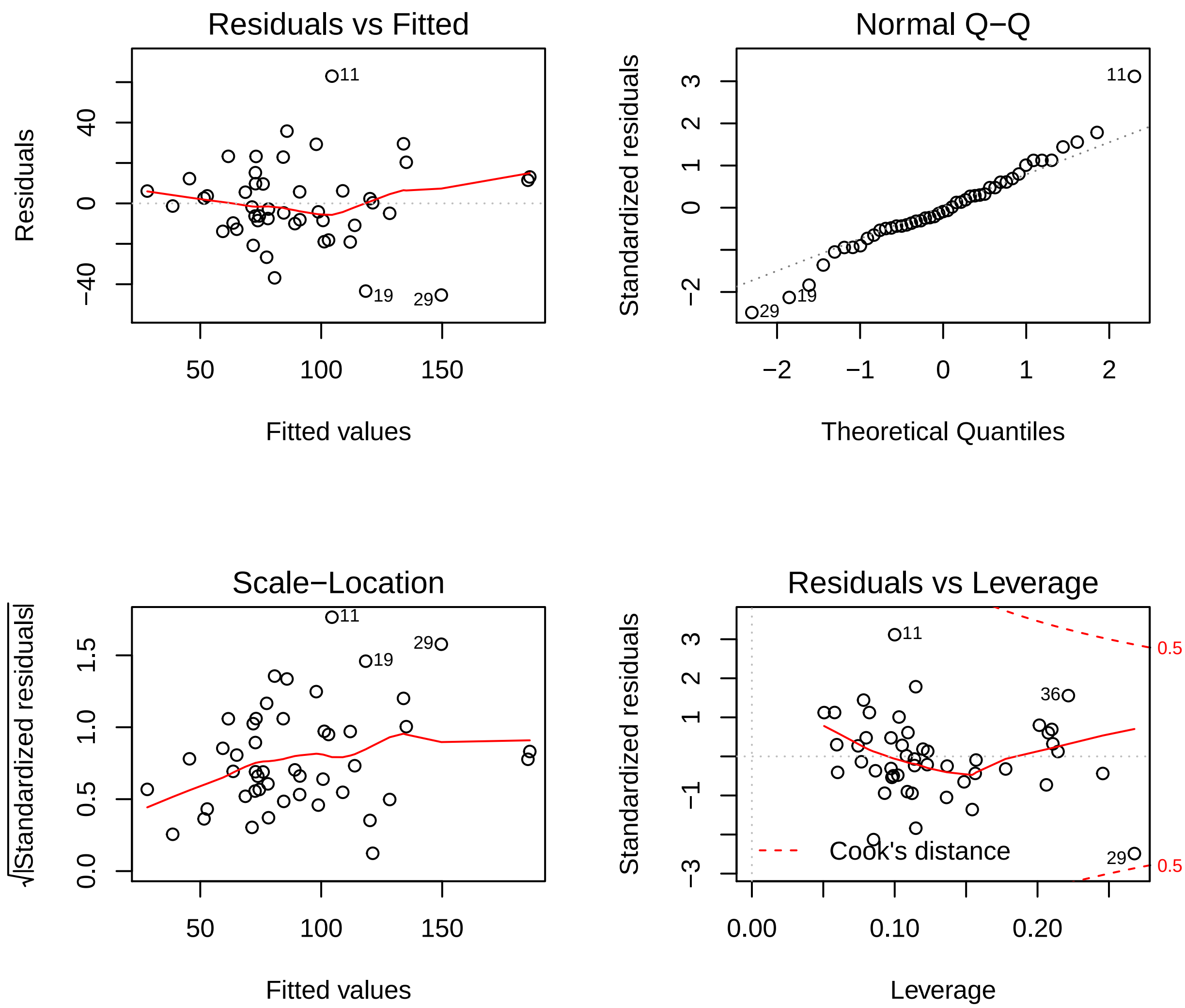
```
##      Min       1Q   Median       3Q      Max 
## -45.344  -9.859  -1.807  10.603  62.964 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -524.3743    95.1156  -5.513 2.13e-06 ***
## Age            1.0198     0.3532   2.887 0.006175 ** 
## Ed             2.0308     0.4742   4.283 0.000109 ***
## Ex0            1.2331     0.1416   8.706 7.26e-11 ***
## U2             0.9136     0.4341   2.105 0.041496 *  
## X              0.6349     0.1468   4.324 9.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 21.3 on 41 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.6967 
## F-statistic: 22.13 on 5 and 41 DF,  p-value: 1.105e-10
```

Now, vif values are all below 4 and all regressors are significant, so this is the minimal model. The equation for the model is

$$R = 1.02 * \texttt{Age} + 3.031 * \texttt{Ed} + 1.233 * \texttt{Ex0} + 0.914 * \texttt{U2} + 0.635 * \texttt{X}$$

(d) Check the validity of the model assumptions starting with diagnostics plots and carry out any tests that are necessary. **Comment all your steps.**

```r
par(mfrow = c(2,2))
plot(modelAIC2)
```




```r
par(mfrow=c(1,1))
```

All the plots look reasonable. In the first plot, the distribution of the residuals looks random and approximately symmetric. The quantile plot shows some departures at the tails, but in general seems reasonable. We can confirm this using the Shapiro-Wilk test on the standardized residuals:

```r
shapiro.test(rstandard(modelAIC2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(modelAIC2)
## W = 0.97494, p-value = 0.403
```

The third plot also looks reasonable although a slight increasing pattern can be seen in the local regression line. To confirm whether this is significant, we use the ncv test

```r
ncvTest(modelAIC2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.758246, Df = 1, p = 0.052548
```

Since the p-value is above 0.05 threshold, we conclude that there is no heteroscedasticity. Finally, the fourth plot shows one point with high leverage and large value for Cook's distance (close to the contour line), which is point 29. This point should be checked in a more thorough study of the regression model.