STAT 210
Applied Statistics and Data Analysis:
Homework 8

Due on Nov. 13/2022

Question 1

For this question we will use the data set `dataB` which has a response variable `res` and five covariates.

(i) Do a exploratory analysis of this data set, including a scatterplot matrix and a graphical representation of the correlation matrix. Comment on your results.

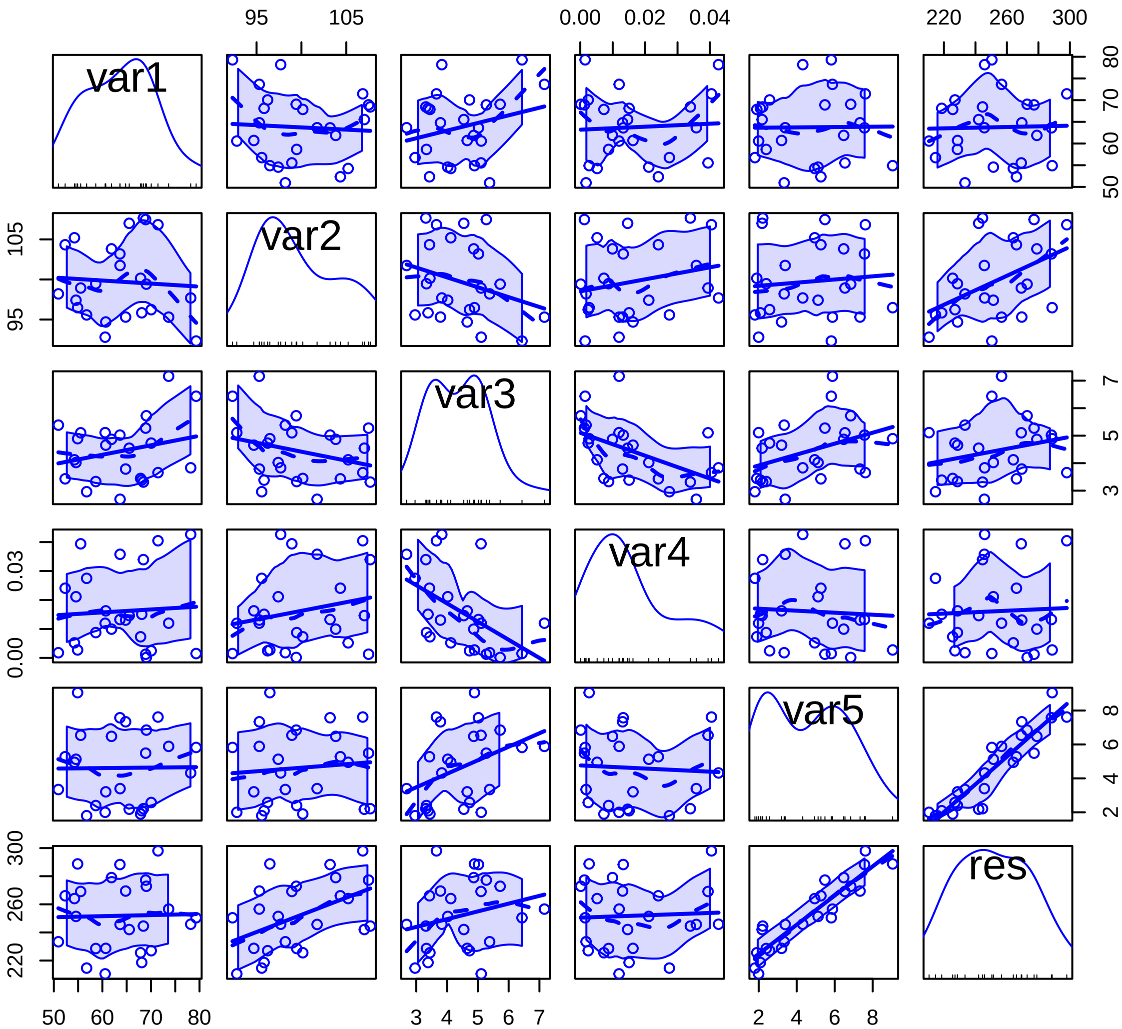
```
library(car)
library(corrplot)
dataB <- read.table('dataB')
str(dataB)
```

```
## 'data.frame':    25 obs. of  6 variables:
##  $ var1: num  68.1 78.2 54.2 54.6 56.8 ...
##  $ var2: num  95.8 97.7 105.2 97.4 95.6 ...
##  $ var3: num  3.38 3.83 4.12 4.02 2.96 ...
##  $ var4: num  0.01506 0.04265 0.00519 0.02113 0.02749 ...
##  $ var5: num  2.09 4.32 4.95 5.13 1.8 ...
##  $ res : num  219 246 264 251 215 ...
```

```
library(psych)
describe(dataB)
```

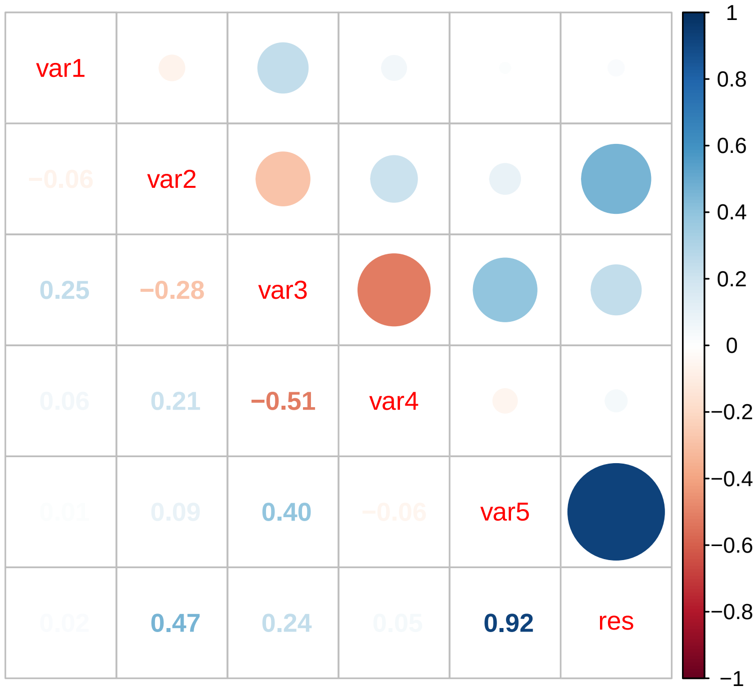
```
##      vars  n   mean    sd median trimmed   mad    min    max range skew
## var1    1 25  63.74  7.84  63.64   63.47  8.01  50.94  79.30 28.36 0.17
## var2    2 25  99.72  4.77  98.93   99.66  5.37  92.32 107.66 15.35 0.30
## var3    3 25   4.43  1.10   4.55    4.36  1.12   2.69   7.16  4.47 0.52
## var4    4 25   0.02  0.01   0.01    0.02  0.02   0.00   0.04  0.04 0.66
## var5    5 25   4.62  2.20   4.95    4.53  2.81   1.80   9.05  7.25 0.22
## res     6 25 251.86 24.82 250.43  251.64 32.26 210.53 298.03 87.50 0.08
##      kurtosis   se
## var1    -0.97 1.57
## var2    -1.28 0.95
## var3    -0.33 0.22
## var4    -0.93 0.00
## var5    -1.33 0.44
## res     -1.19 4.96
```

```
scatterplotMatrix(dataB)
```



Looking at the graphs in the bottom row, where `res` is in the y-axis, `var1` and `var4` seem to have no relation with `res`, while `var5` shows a strong linear relation with positive slope. The remaining variables, `var2` and `var3` have a (moderate) linear relation with `res` with positive slope, but there is more variability in these cases.

```
dataB.cor <- cor(dataB)
corrplot.mixed(dataB.cor)
```



Variables `var3` and `var4` have a moderately large negative correlation that may cause multicollinearity problems. Variable `var5` has an important positive correlation with `res`. This was commented in the previous graph. Variables `var2` and `var3` have an moderate positive correlation with `res`.

(ii) Fit a complete model for `res` including all the other variables. Produce a summary table and interpret the t tests in the table. What is the p-value for the overall significance test for the regression?

```
lm1 <- lm(res ~ var1+var2+var3+var4+var5, data = dataB)
summary(lm1)
```

```
##
## Call:
## lm(formula = res ~ var1 + var2 + var3 + var4 + var5, data = dataB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2380 -0.6552 -0.0618  0.3911  2.0899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.17184    4.66622  -0.894 0.382480
## var1         0.12348    0.02562   4.820 0.000119 ***
## var2         2.02474    0.04227  47.898  < 2e-16 ***
## var3        -0.04097    0.24458  -0.168 0.868727
## var4        18.32082   17.01703   1.077 0.295132
## var5         9.99242    0.09817 101.787  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.92 on 19 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic:  3488 on 5 and 19 DF,  p-value: < 2.2e-16
```

Variables `var3` and `var4` have large p-values and the coeffcients are not significantly different form zero. The other variables have small p-values. The p-value for the overall significance test is 0 (< 2.2e-16) and appears at the bottom of the summary table.

(iii) Starting with the model fitted in section (ii), fit a minimal model using a backwards selection procedure with a critical α of 0.15.

We drop `var3`, since it has the largest p-value.

```
lm2 <- update(lm1, .~.-var3)
summary(lm2)
```

```
##
## Call:
## lm(formula = res ~ var1 + var2 + var4 + var5, data = dataB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23126 -0.68464 -0.07108  0.40683  2.09869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.44092    4.27329  -1.039    0.311
## var1         0.12201    0.02347   5.199 4.36e-05 ***
## var2         2.02667    0.03966  51.100  < 2e-16 ***
## var4        19.85867   13.97615   1.421    0.171
## var5         9.98449    0.08390 119.000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8974 on 20 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9987
## F-statistic:  4583 on 4 and 20 DF,  p-value: < 2.2e-16
```

`var4` has a large p-value and so we drop it from the model.

```
lm3 <- update(lm2, .~. - var4)
summary(lm3)
```

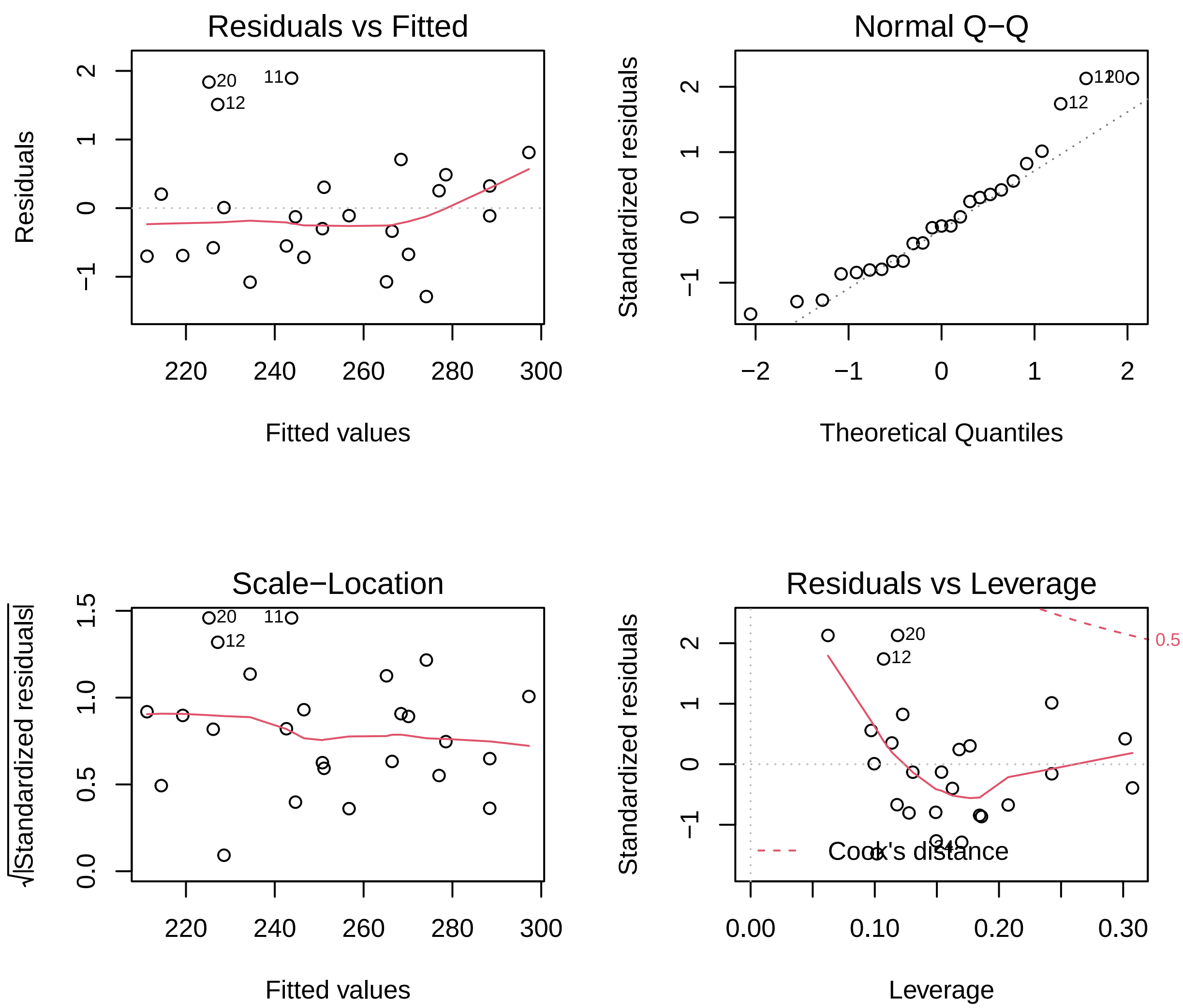
```
##
## Call:
## lm(formula = res ~ var1 + var2 + var5, data = dataB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2885 -0.6737 -0.1126  0.3231  1.8934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.49173    4.30971  -1.274    0.216
## var1         0.12454    0.02396   5.197 3.77e-05 ***
## var2         2.03923    0.03959  51.509  < 2e-16 ***
## var5         9.97503    0.08564 116.471  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9189 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  5827 on 3 and 21 DF,  p-value: < 2.2e-16
```

The minimal adequate model includes `var1`, `var2` and `var5`.

(iv) Plot the standard diagnostic graphs for the model that you selected and comment on what you observe.

```
par(mfrow=c(2,2))
plot(lm3)
```




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

All the plots look reasonable in this case. We test for normality and homoscedasticity:

```
shapiro.test(rstandard(lm3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(lm3)
## W = 0.93522, p-value = 0.1147
```

```
ncvTest(lm3)
```

```
## Non-constant Variance Score Test
```

```
## Variance formula: ~ fitted.values
## Chisquare = 1.214087, Df = 1, p = 0.27052
```

Both tests have large p-values and the null hypotheses are not rejected.

(v) Predict the `res` value for a subject with covariates `(var1,var2,var3,var4,var5) = (65,100,50,0.02,3)`. Add a confidence interval at level 98%.

`(var1,var2,var3,var4,var5) = (65,100,50,0.02,3)`.

```
newdata = data.frame(var1 = 65, var2 = 100, var5 = 3)
predict(lm3,newdata, level = 0.98, interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 236.4518 235.8644 237.0392
```

(vi) Print an anova table for the final model and find the estimated variance of the errors. Describe explicitly the sampling distribution for the estimated parameters.

```
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: res
##           Df  Sum Sq Mean Sq   F value    Pr(>F)
## var1       1     8.6     8.6    10.194  0.004377 **
## var2       1  3297.5  3297.5  3905.042 < 2.2e-16 ***
## var5       1 11455.1 11455.1 13565.498 < 2.2e-16 ***
## Residuals 21    17.7     0.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated variance for the errors is the mean square error, which is 0.8. The sampling distribution for the estimated parameters is normal

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_5)' \sim N(\boldsymbol{\beta}, \mathbf{V})$$

where the covariance matrix $\mathbf{V} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is

```
vcov(lm3)
```

```
##               (Intercept)          var1          var2          var5
## (Intercept) 18.5735748406 -4.249603e-02 -1.587225e-01 -6.146995e-04
## var1        -0.0424960346  5.742577e-04  6.062873e-05 -3.357062e-05
## var2        -0.1587224922  6.062873e-05  1.567337e-03 -3.119544e-04
## var5        -0.0006146995 -3.357062e-05 -3.119544e-04  7.334879e-03
```

Question 2

The file `dataC` has information on two variables, `yvar` and `xvar`. We want to build a regression model for `yvar` as a function of `xvar`.

(i) Fit a simple regression model for `yvar` in terms of `xvar`. Print the summary table and comment on the results. Draw a scatterplot and add the regression line. Comment.

```
dataC <- read.table('dataC', header = T)
str(dataC)
```

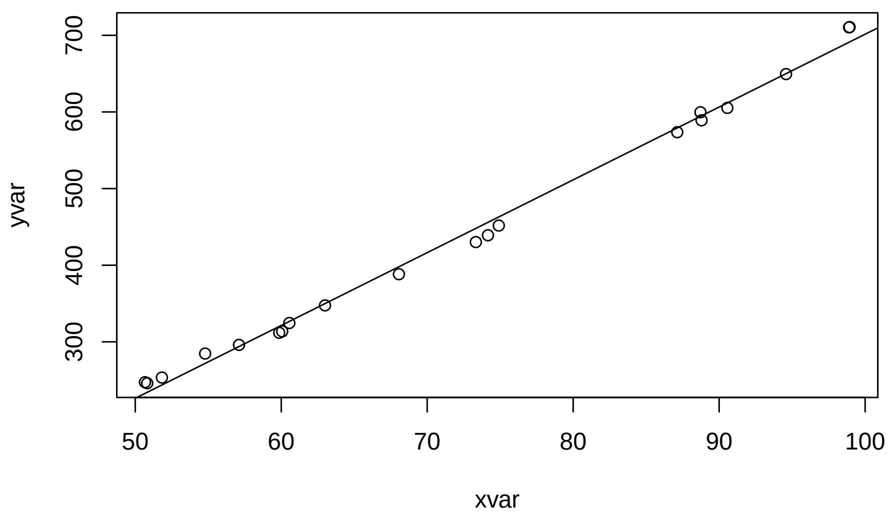
```
## 'data.frame':    20 obs. of  2 variables:
```

```
##  $ yvar: num  600 312 710 314 388 ...
##  $ xvar: num  88.7 59.9 98.9 60.1 68.1 ...
```

```r
library(car)
mod1 <- lm(yvar ~ xvar, data = dataC)
summary(mod1)
```

```
##
## Call:
## lm(formula = yvar ~ xvar, data = dataC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.883  -8.306  -2.206  10.221  19.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -248.562     11.799  -21.07 3.92e-14 ***
## xvar           9.499      0.159   59.74  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.74 on 18 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9947
## F-statistic:  3569 on 1 and 18 DF,  p-value: < 2.2e-16
```

```r
plot(yvar ~ xvar, data = dataC)
abline(mod1)
```

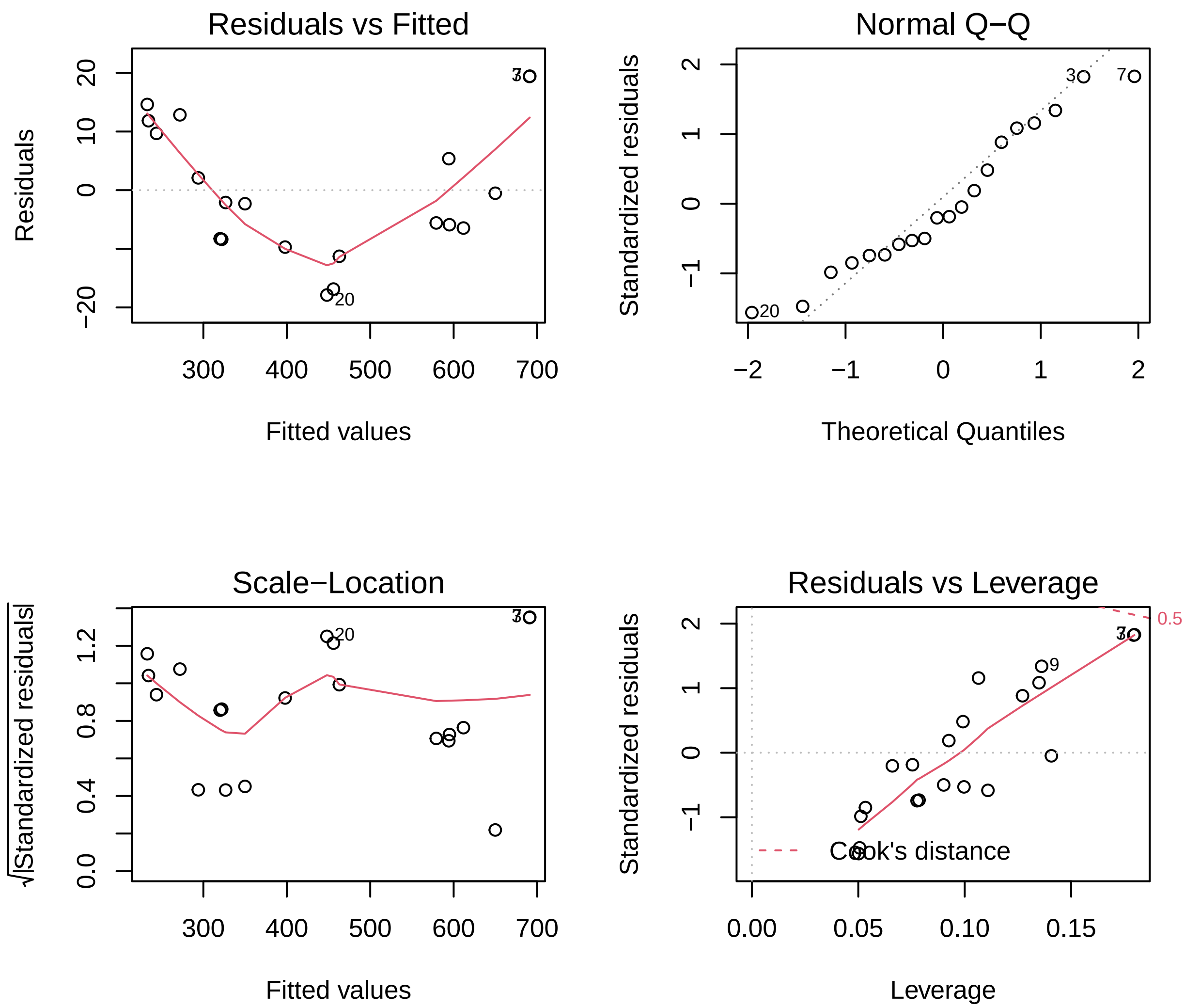


The fit looks good but the central points are mostly below the regression line, while at the extremes they are above. This shows that the data has a curvature that the model is not capturing.

(ii) State clearly the assumptions on which the model is based and, using the standard diagnostic plots and any tests that are necessary, verify if these assumptions are valid in this case.

The errors are assumed to be independent, having a normal distribution with mean zero and common variance σ^2.

```r
par(mfrow=c(2,2))
plot(mod1)
```




```
par(mfrow=c(1,1))
shapiro.test(rstandard(mod1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod1)
## W = 0.94271, p-value = 0.2697
```

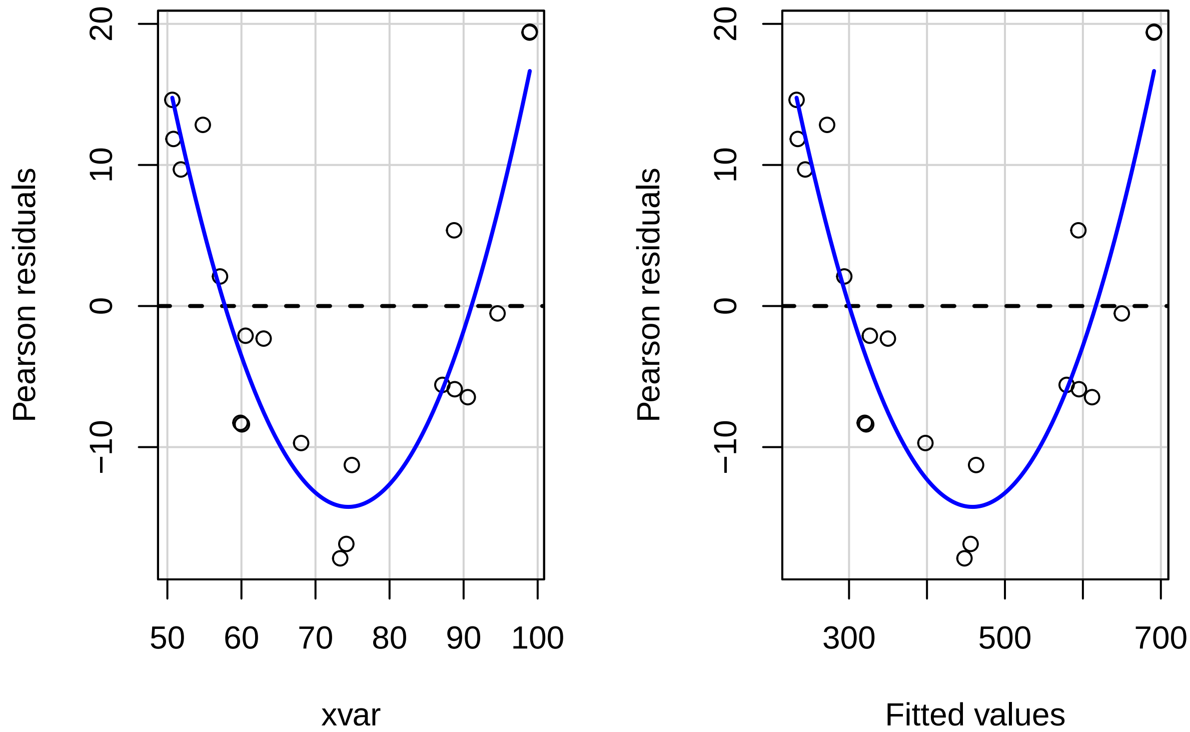
```
ncvTest(mod1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4522966, Df = 1, p = 0.50125
```

The residuals vs fitted plot shows a quadratic pattern and the residuals are not symmetrically distributed. The model is not adequate.

(iii) Use the function `residualPlots` in the package `car`. This function was introduced in problem 2 of Problem List 8. The result of applying this function is twofold. On the one hand, graphs of residuals against fitted values and regressors are plotted, including in blue a quadratic term, and on the other hand, a couple of tests are performed and printed in the console. The first test tests whether a quadratic term in the regressor variable would be significant. Interpret the result that you get.

```r
residualPlots(mod1)
```




```
##            Test stat Pr(>|Test stat|)
## xvar          9.8964        1.803e-08 ***
## Tukey test    9.8964        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plots and tests indicate that a quadratic term in `xvar` should be included in the model.

(iv) Add a quadratic term to the initial regression model. Print the summary table, and interpret the results. Draw the diagnostic plots and comment on them.

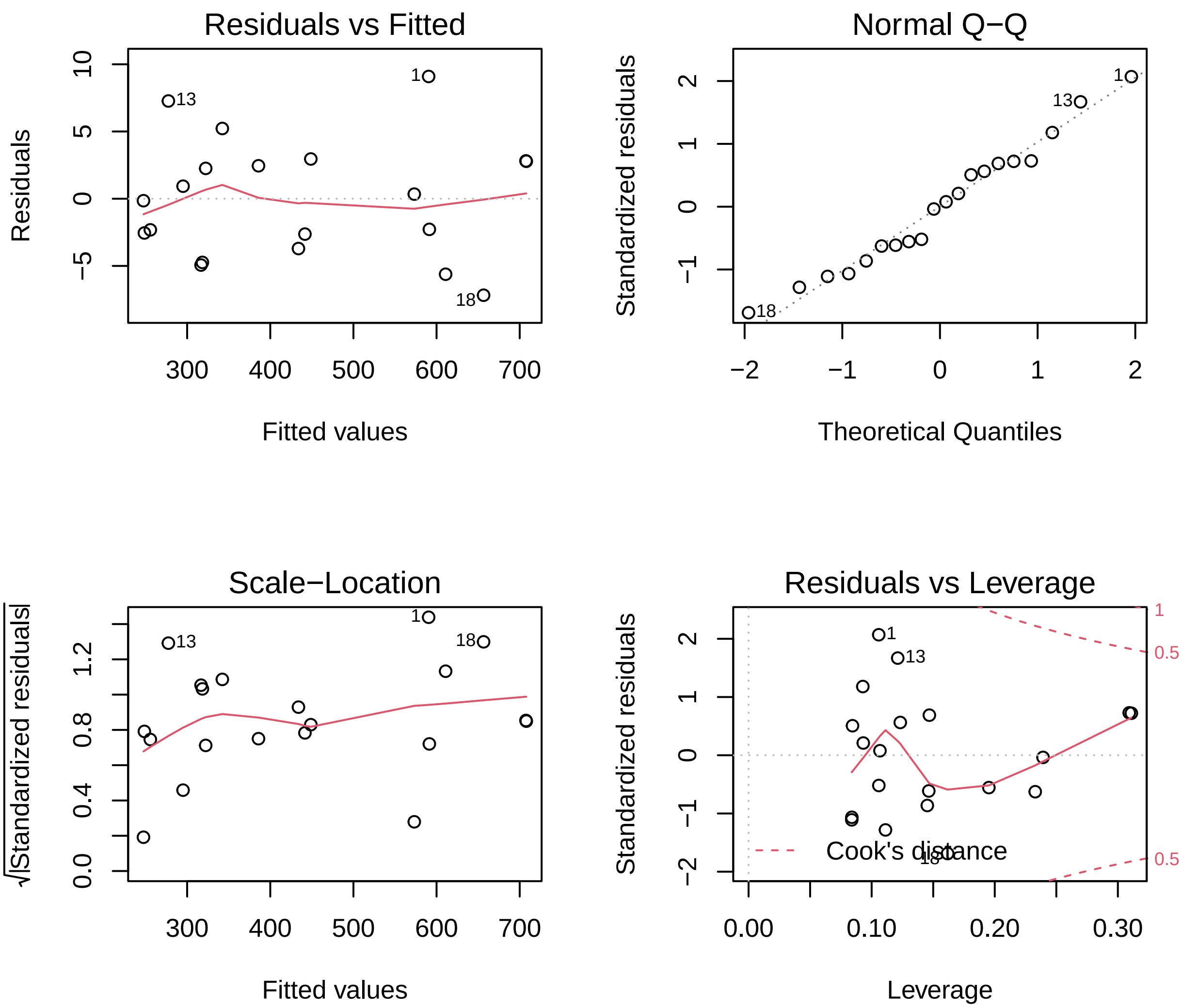
```r
mod2 <-  lm(yvar ~ xvar + I(xvar^2), data = dataC)
summary(mod2)
```

```
##
## Call:
## lm(formula = yvar ~ xvar + I(xvar^2), data = dataC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1816 -2.9013  0.0967  2.7915  9.0942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.862962  27.721554   0.789    0.441
## xvar         1.849283   0.775584   2.384    0.029 *
## I(xvar^2)    0.051399   0.005194   9.896  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.647 on 17 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9992
## F-statistic: 1.144e+04 on 2 and 17 DF,  p-value: < 2.2e-16
```

The quadratic term is highly significant. The R^2 is almost one.

```
par(mfrow=c(2,2))
plot(mod2)
```




```
par(mfrow=c(1,1))
shapiro.test(rstandard(mod2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod2)
## W = 0.97071, p-value = 0.7697
```

```
ncvTest(mod2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6098989, Df = 1, p = 0.43483
```

The plots look better now and the tests are consistent with the hypotheses.

(v) Write an equation for the model. Do a scatter plot and add the initial regression line and the curve for the quadratic model that you fitted in (v).

The equation for the model is

$$yvar = 21.863 + 1.85 * xvar + 0.0514 * (xvar)^2$$

```
plot(yvar ~ xvar, data = dataC)
abline(mod1)
curve(21.863 + 1.85*x + 0.0514*x^2,50,100, add = T, col = 'red')
```