STAT 210
Applied Statistics and Data Analysis:
Homework 9 - Solution

Due on Nov. 20/2022

Question 1

Consider the `prostate` data set in the `faraway` package. Consider `lpsa` as the response variable and exclude the variables `svi` and `gleason` from the analysis.

```
library(faraway)
str(prostate)
```

```
## 'data.frame':    97 obs. of  9 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

```
prost <- prostate[, -c(5,7)]
str(prost)
```

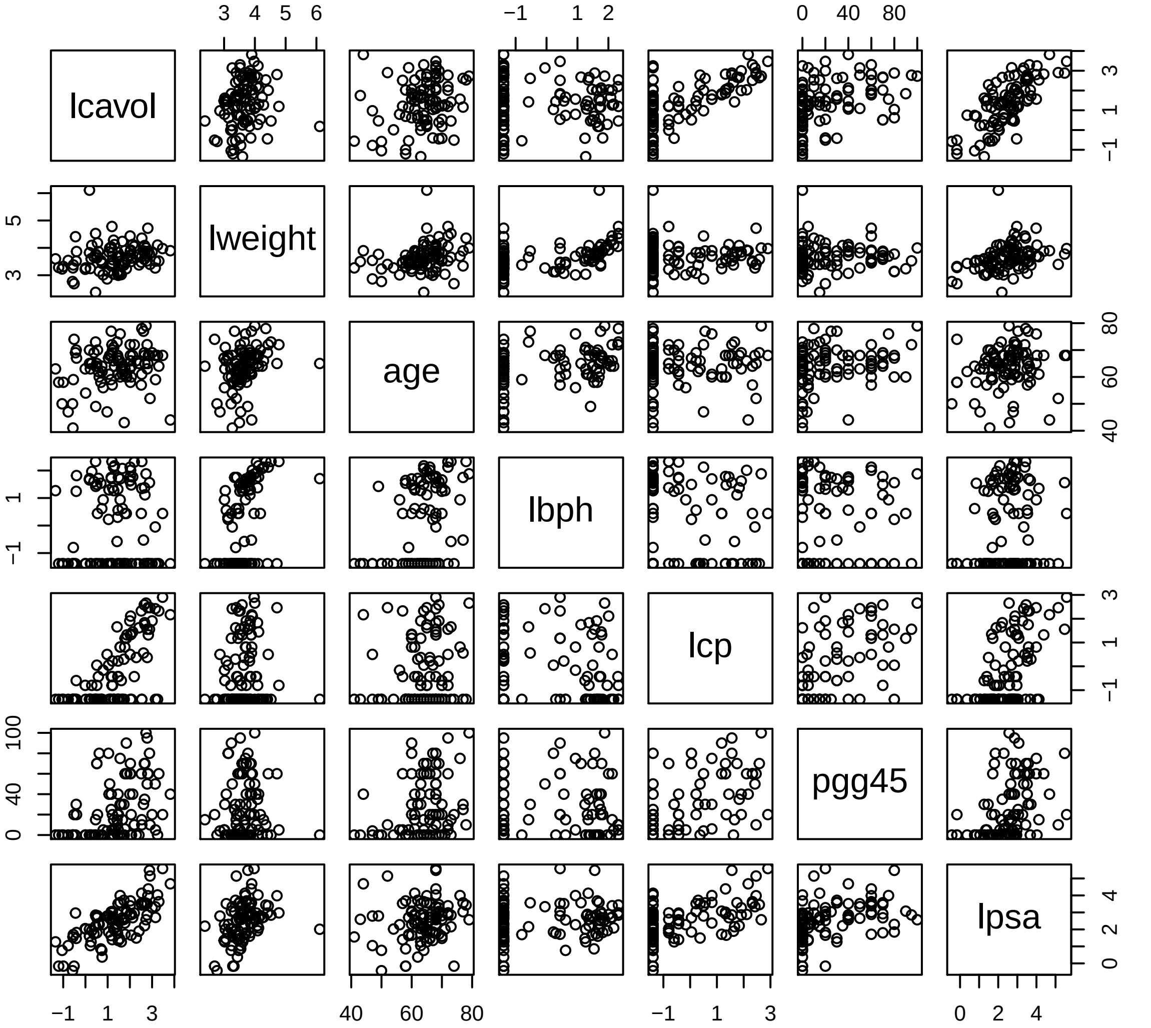
```
## 'data.frame':    97 obs. of  7 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

(a) Do an exploratory analysis of the data. Do a matrix of plots. Which variables seem to have a linear relationship with the response? Compute and plot the correlation coefficients for the regressors. Comment on what you obtain.

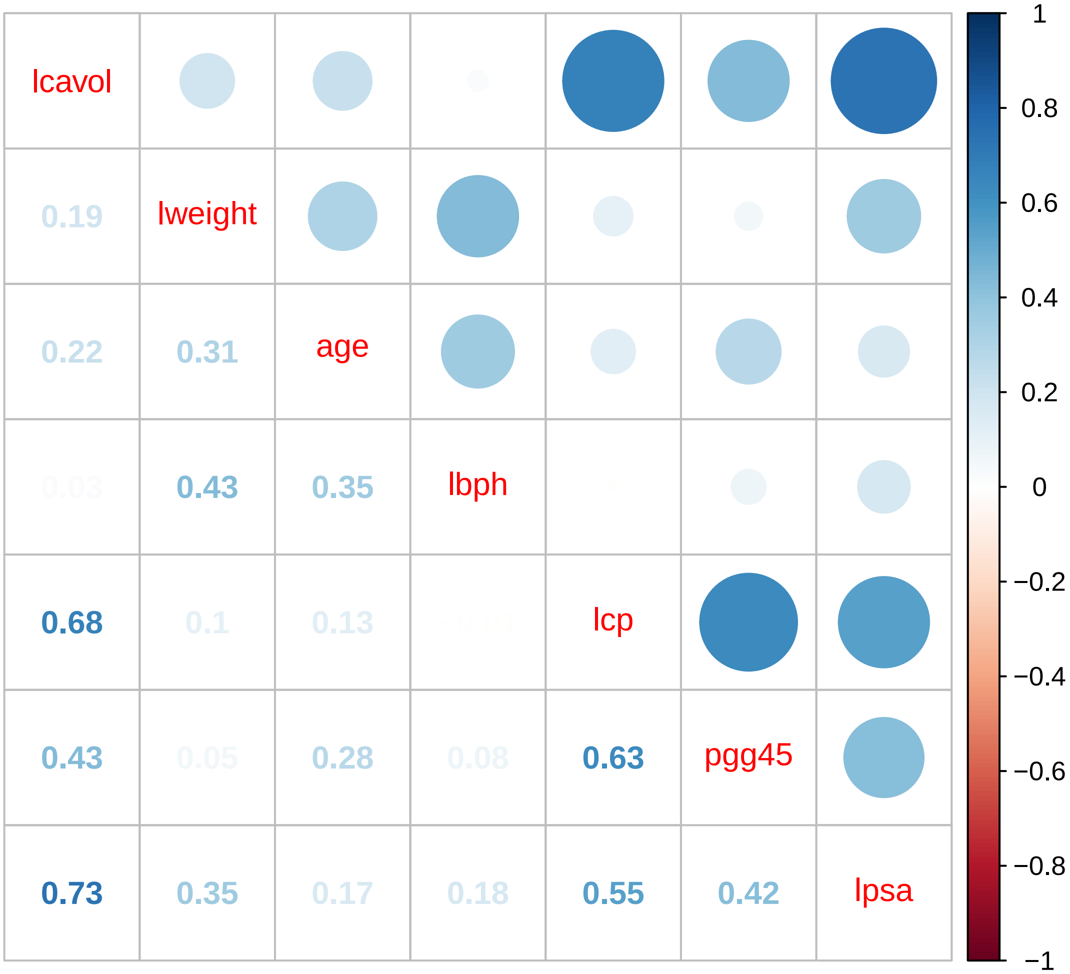
```
options(width = 90)
library(psych)
describe(prost)
```

```
##         vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
## lcavol     1 97  1.35 1.18   1.45    1.39 1.28 -1.35  3.82  5.17 -0.24    -0.60 0.12
## lweight    2 97  3.65 0.50   3.62    3.63 0.38  2.37  6.11  3.73  1.18     5.02 0.05
## age        3 97 63.87 7.45  65.00   64.47 5.93 41.00 79.00 38.00 -0.80     0.96 0.76
## lbph       4 97  0.10 1.45   0.30    0.03 2.50 -1.39  2.33  3.71  0.13    -1.75 0.15
## lcp        5 97 -0.18 1.40  -0.80   -0.34 0.87 -1.39  2.90  4.29  0.71    -1.01 0.14
## pgg45       6 97 24.38 28.20  15.00   20.57 22.24  0.00 100.00 100.00  0.94    -0.37 2.86
## lpsa        7 97  2.48  1.15   2.59    2.48  1.15 -0.43   5.58   6.01  0.00     0.43 0.12
```

```
plot(prost)
```



```
library(corrplot)
prost.cor <- cor(prost)
corrplot.mixed(prost.cor)
```



The variable that seems to have a good linear relation with `lpsa` is `lcavol`; `lcp` and `pgg45` also have a moderate linear relation with `lpsa`. On the other hand, `lcp` and `pgg45` and `lcp` and `lcavol` have a correlation over 0.6 and may have high collinearity.

(b) Fit a model for `lpsa` with all the other variables as predictors. Calculate the variance inflation factors and eliminate variables with vif greater than two.

```
mod1 <- lm(lpsa ~ ., data = prost)
summary(mod1)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41989 -0.44616 -0.01308  0.40366  1.76377
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.833707   0.868392   0.960  0.33960
## lcavol       0.628344   0.089299   7.036 3.75e-10 ***
## lweight      0.496270   0.175106   2.834  0.00567 **
## age         -0.018302   0.011593  -1.579  0.11791
## lbph         0.078998   0.060145   1.313  0.19237
## lcp          0.013881   0.086208   0.161  0.87244
## pgg45        0.006034   0.003590   1.681  0.09625 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7386 on 90 degrees of freedom
## Multiple R-squared:  0.6161, Adjusted R-squared:  0.5905
## F-statistic: 24.08 on 6 and 90 DF,  p-value: < 2.2e-16
```

```
vif(mod1)
```

```
##   lcavol  lweight      age     lbph      lcp    pgg45
## 1.949185 1.330687 1.310768 1.339754 2.556673 1.803585
```

Only two regressors appear to be significant. One variable, `lcp`, has a large vif and must be eliminated from the model.

```
mod2 <- update(mod1, .~. - lcp)
vif(mod2)
```

```
##   lcavol  lweight      age     lbph    pgg45
## 1.303767 1.329029 1.282407 1.337491 1.300106
```

All vif values are now below 2. We print the summary table for this model

```
summary(mod2)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + pgg45, data = prost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40484 -0.44324 -0.02214  0.40586  1.78837
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.826541   0.862597   0.958   0.3405
## lcavol       0.636618   0.072641   8.764 9.85e-14 ***
## lweight      0.497266   0.174058   2.857   0.0053 **
## age         -0.018576   0.011405  -1.629   0.1068
## lbph         0.078599   0.059772   1.315   0.1918
## pgg45        0.006339   0.003031   2.091   0.0393 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7347 on 91 degrees of freedom
## Multiple R-squared:  0.616,  Adjusted R-squared:  0.5949
## F-statistic:  29.2 on 5 and 91 DF,  p-value: < 2.2e-16
```

(c) Starting with the variables selected in (b), do a variable selection procedure using backward elimination with a p to remove equal to 0.15. Do also variable selection using the BIC criterion. Compare the models that you get. Do residual analysis for both of them. Comment on your results.

We remove `lbph`

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

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + pgg45, data = prost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47277 -0.44626 -0.01144  0.44526  1.75982
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.284167   0.760596   0.374 0.709553
## lcavol       0.624916   0.072379   8.634 1.71e-13 ***
## lweight      0.584455   0.161571   3.617 0.000486 ***
## age         -0.014766   0.011075  -1.333 0.185716
## pgg45        0.006513   0.003040   2.142 0.034834 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7376 on 92 degrees of freedom
## Multiple R-squared:  0.6087, Adjusted R-squared:  0.5917
## F-statistic: 35.78 on 4 and 92 DF,  p-value: < 2.2e-16
```

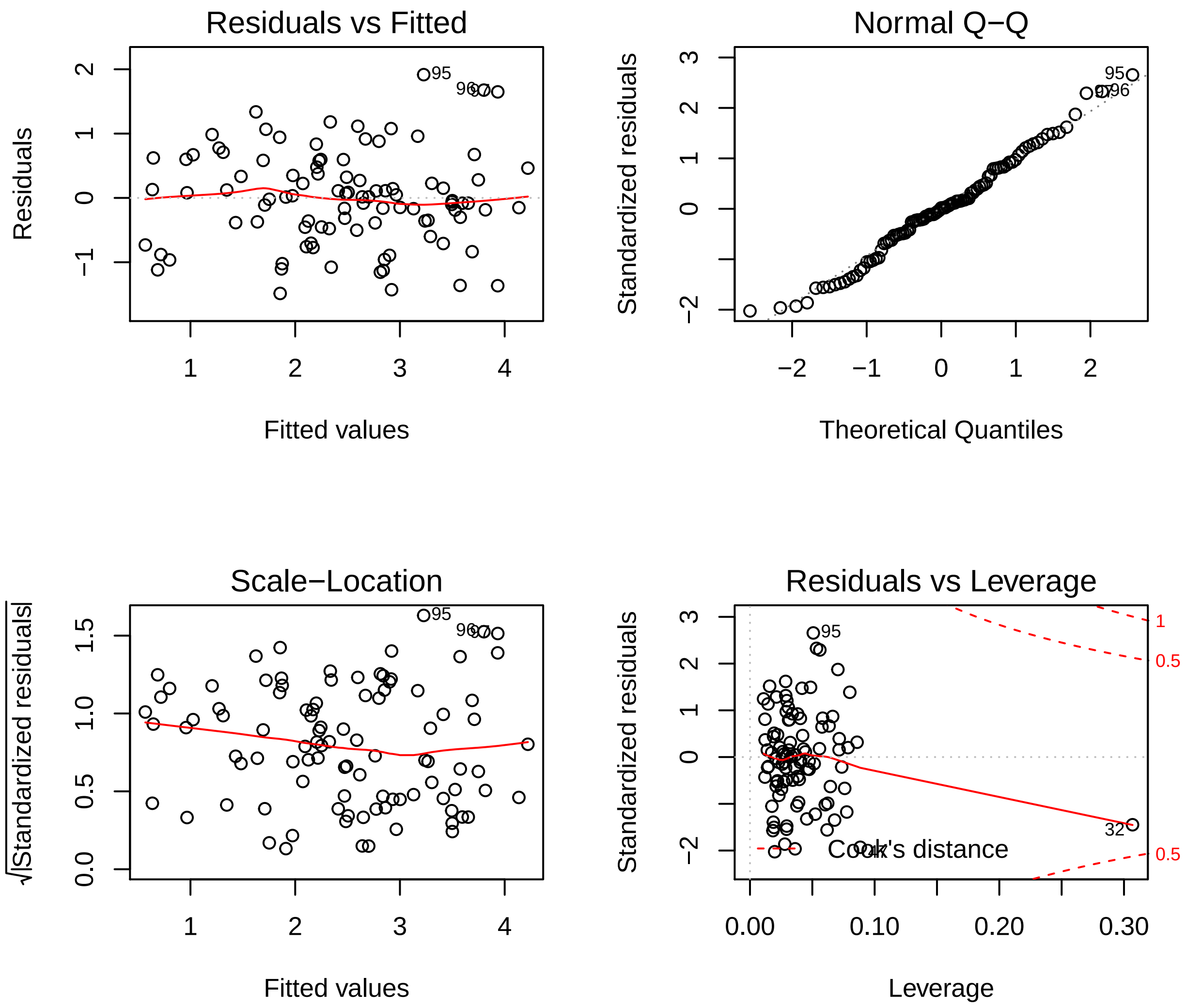
We remove `age`

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

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + pgg45, data = prost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48488 -0.45153  0.01277  0.47780  1.91618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.399910   0.563837  -0.709  0.47993
## lcavol       0.618402   0.072516   8.528 2.65e-13 ***
## lweight      0.521988   0.155275   3.362  0.00113 **
## pgg45        0.005610   0.002977   1.885  0.06257 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7407 on 93 degrees of freedom
## Multiple R-squared:  0.6012, Adjusted R-squared:  0.5883
## F-statistic: 46.73 on 3 and 93 DF,  p-value: < 2.2e-16
```

The final model with this procedure retains as regressors the variables `lcavol`, `lweight` and `pgg45`.

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



We also fit a model using BIC.

```
library(MASS)
stepAIC(mod2, k = log(94))
```

```
## Start:  AIC=-38.75
## lpsa ~ lcavol + lweight + age + lbph + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lbph     1     0.933 50.050 -41.467
## - age      1     1.432 50.549 -40.505
## <none>                 49.117 -38.750
## - pgg45    1     2.360 51.477 -38.740
## - lweight  1     4.405 53.522 -34.961
## - lcavol   1    41.455 90.572  16.066
##
## Step:  AIC=-41.47
## lpsa ~ lcavol + lweight + age + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - age      1     0.967 51.017 -44.154
## <none>                 50.050 -41.467
## - pgg45    1     2.496 52.546 -41.289
## - lweight  1     7.119 57.169 -33.111
## - lcavol   1    40.554 90.604  11.557
##
## Step:  AIC=-44.15
## lpsa ~ lcavol + lweight + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1     1.949 52.966 -45.061
## <none>                 51.017 -44.154
## - lweight  1     6.199 57.217 -37.573
## - lcavol   1    39.895 90.912   7.342
##
## Step:  AIC=-45.06
## lpsa ~ lcavol + lweight
##
##           Df Sum of Sq     RSS     AIC
## <none>                  52.966 -45.061
## - lweight  1     5.949  58.915 -39.279
## - lcavol   1    58.910 111.876  22.927
```

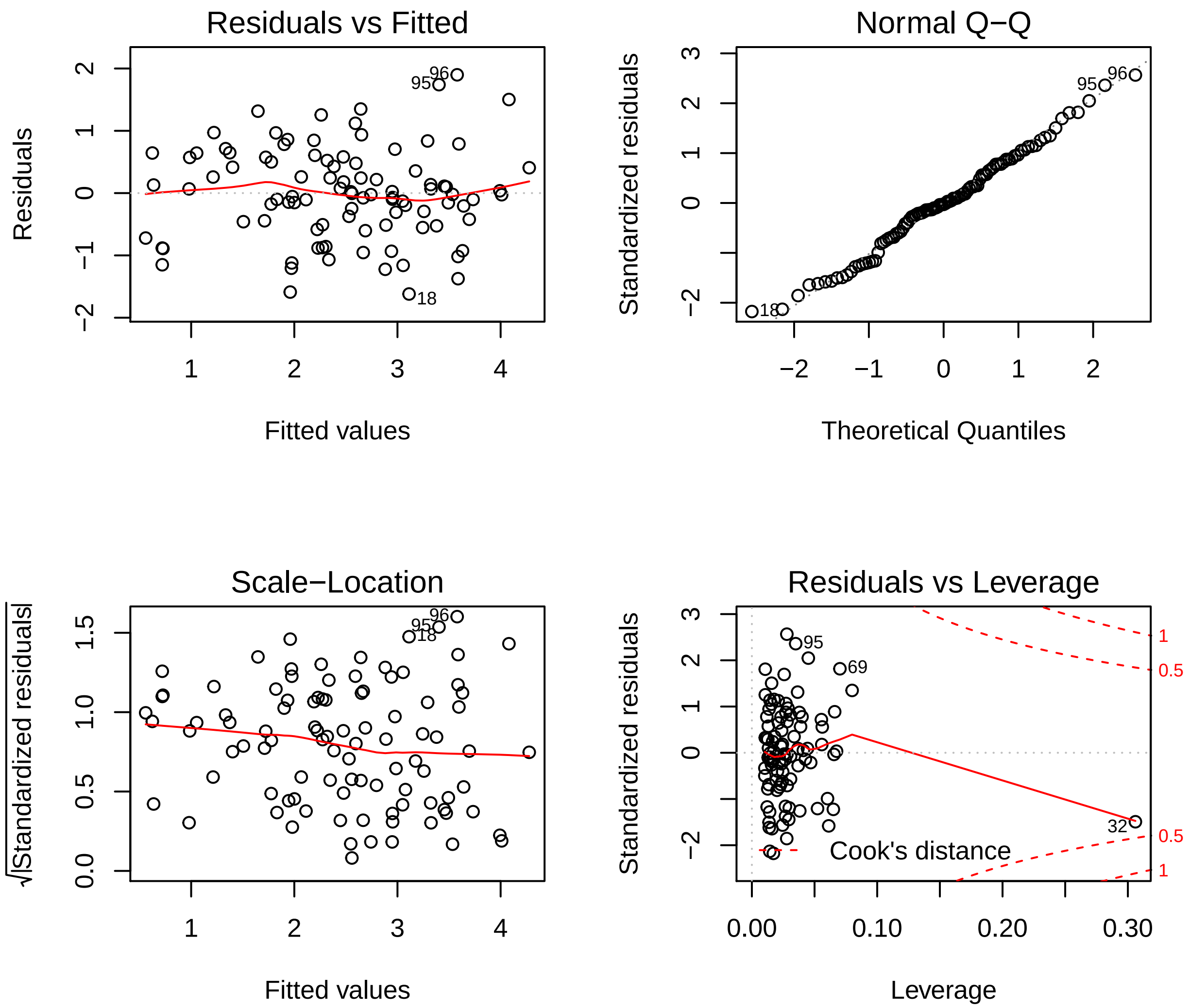
```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight, data = prost)
##
## Coefficients:
## (Intercept)       lcavol      lweight
##     -0.3026       0.6775       0.5109
```

This procedures selects a model with only two regressors, `lcavol` and `lweight`.

```
mod5 <- update(mod4, .~. - pgg45)
summary(mod5)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight, data = prost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61965 -0.50778 -0.02095  0.52291  1.89885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.30262    0.56904  -0.532  0.59612
## lcavol       0.67753    0.06626  10.225  < 2e-16 ***
## lweight      0.51095    0.15726   3.249  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7506 on 94 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5771
## F-statistic: 66.51 on 2 and 94 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod5)
```



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

(d) Which model would you select and why?

This is not a clear-cut decision, but using the principle of parsimony, I would choose the simpler model. Here are my reasons:

We start with the diagnostic plots to see if there are important differences that point to one or both models not satisfying the assumptions used to build the model. In this case, the diagnostic plots are very similar, and none shows important deficiencies in the model. Next, the p-value for the additional term in `mod4` is moderately large. We have retained this term because the p-value is smaller than the threshold of 0.15 that the question set. Third, the adjusted R^2 favors `mod4`, but the difference between the R^2s for both models is not big, about 1%. Finally, the simpler model has better BIC.

Therefore, I would conclude that the two models are equivalent and we should keep the simpler one, which is `mod5`.

The equation for the model is

$$\text{lpsa} = -0.30262 + 0.67753 \times \text{lcavol} + 0.51095 \times \text{lweight}$$

(e) Suppose a new patient with the following values arrives:

Table 1: Variables for a new patient

lcavol	lweight	age	Ibph	lcp	pgg45
1.44692	3.623	65	0.30	-0.799	15.0

Predict the `lpsa` for this patient along with appropriate 98% prediction and confidence intervals.

Confidence interval:

```r
new.data <- data.frame(lcavol = 1.4469, lweight = 3.623)
predict(mod5, new.data, interval = 'c')
```

```
##        fit      lwr      upr
## 1 2.528863 2.376564 2.681163
```

Prediction interval:

```r
predict(mod5, new.data, interval = 'p')
```

```
##        fit      lwr      upr
## 1 2.528863 1.030675 4.027052
```

Question 2

For this question use the data set `Birthweight.csv`. We will consider only the variables `birthwt`, `mppwt` and `smoker`. They represent the weight of the baby at birth, the weight of the mother before pregnancy and whether the mother smokes, with 1 indicating that the mother is a smoker.

```r
data1 <- read.csv('Birthweight.csv')
str(data1)
```

```
## 'data.frame':    42 obs. of  10 variables:
##  $ id               : int  1313 431 808 300 516 321 1363 575 822 1081 ...
##  $ headcirumference: int  12 12 13 12 13 13 12 12 13 14 ...
##  $ length           : int  17 19 19 18 18 19 19 19 19 21 ...
##  $ birthwt          : num  5.8 4.2 6.4 4.5 5.8 6.8 5.2 6.1 7.5 8 ...
##  $ gestation        : int  33 33 34 35 35 37 37 37 38 38 ...
##  $ smoker           : int  0 1 0 1 1 0 1 1 0 0 ...
##  $ motherage        : int  24 20 26 41 20 28 20 19 20 18 ...
##  $ mheight          : int  58 63 65 65 67 62 64 65 62 67 ...
##  $ mppwt            : int  99 109 140 125 125 118 104 132 103 109 ...
##  $ LowBirthWeight   : Factor w/ 2 levels "Low","Normal": 1 1 2 1 1 2 1 2 2 2 ...
```

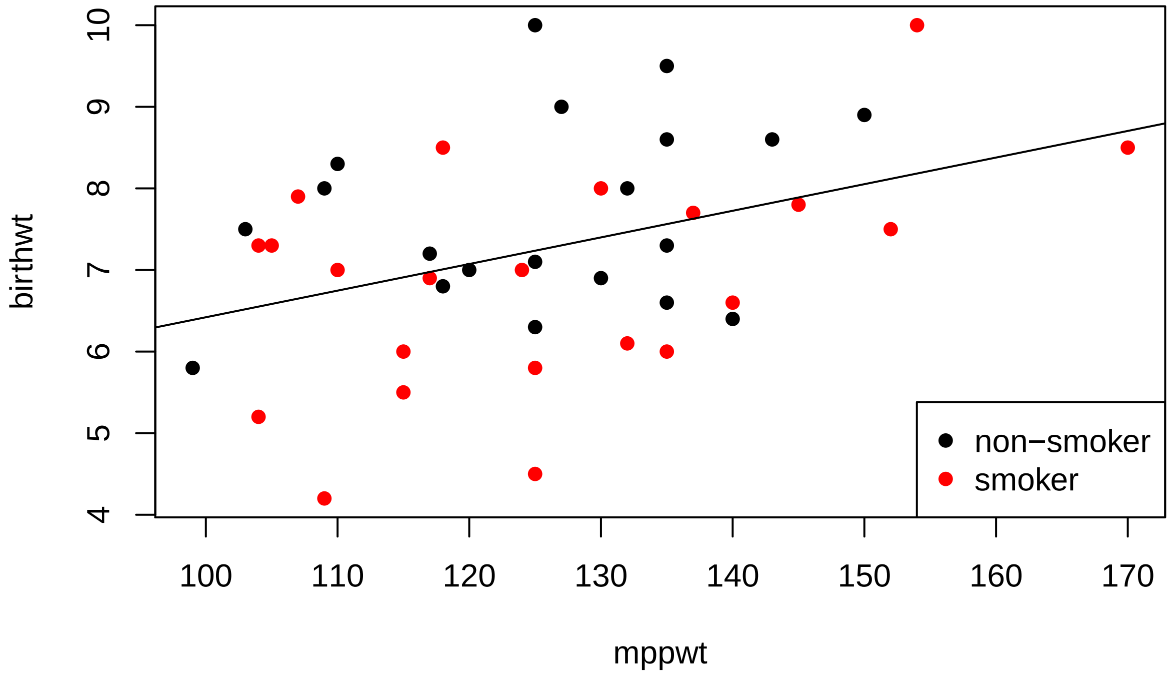
(i) Subset the data corresponding to the variables mentioned above. Plot `birthwt` against `mppwt` and color the dots according to the value of `smoker`. Add a regression line for `birthwt` against `mppwt`. Comment. Print the summary table for the regression and interpret the results.

```r
bwt <- subset(data1, select = c(birthwt,mppwt,smoker))
str(bwt)
```

```
## 'data.frame':    42 obs. of  3 variables:
##  $ birthwt: num  5.8 4.2 6.4 4.5 5.8 6.8 5.2 6.1 7.5 8 ...
##  $ mppwt  : int  99 109 140 125 125 118 104 132 103 109 ...
##  $ smoker : int  0 1 0 1 1 0 1 1 0 0 ...
```

Plot:

```r
plot(birthwt ~ mppwt, col = smoker+1, pch = 16, data = bwt)
mod1 <- lm(birthwt ~ mppwt, data = bwt)
abline(mod1)
legend('bottomright', c('non-smoker','smoker'), pch=c(16,16), col = 1:3)
```



The weight for the non-smokers seems to be bigger than for smokers.

(ii) We want to add `smoker` as a categorical regressor to the previous model. Fit a complete model including interaction and work your way to a minimal adequate model. Write down the equation for you final model and interpret the coefficients.

```
summary(mod1)
```

```
##
## Call:
## lm(formula = birthwt ~ mppwt, data = bwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73632 -0.92977 -0.08206  0.83021  2.76368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.15855    1.54657   2.042   0.0478 *
## mppwt        0.03262    0.01219   2.675   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.24 on 40 degrees of freedom
## Multiple R-squared:  0.1518, Adjusted R-squared:  0.1306
## F-statistic: 7.157 on 1 and 40 DF,  p-value: 0.01077
```

```
mod2 <- lm(birthwt ~ mppwt*smoker, data = bwt)
summary(mod2)
```

```
##
## Call:
## lm(formula = birthwt ~ mppwt * smoker, data = bwt)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.33988 -0.90366  0.09374  0.74718  2.32772 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.263632   2.578573   1.653    0.106
## mppwt         0.027269   0.020412   1.336    0.190
## smoker       -1.894546   3.160700  -0.599    0.552
## mppwt:smoker  0.008497   0.024957   0.340    0.735
## 
## Residual standard error: 1.194 on 38 degrees of freedom
## Multiple R-squared:  0.2526, Adjusted R-squared:  0.1936 
## F-statistic:  4.28 on 3 and 38 DF,  p-value: 0.01068
```

In this model none of the regressors seems to be significant. We start by removing the interaction term.

```r
mod3 <- lm(birthwt ~ mppwt + smoker, data = bwt)
summary(mod3)
```

```
## 
## Call:
## lm(formula = birthwt ~ mppwt + smoker, data = bwt)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.3428 -0.9464  0.1375  0.8069  2.3314 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  3.54944    1.48259   2.394  0.02157 * 
## mppwt        0.03295    0.01161   2.838  0.00717 **
## smoker      -0.82576    0.36476  -2.264  0.02922 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.181 on 39 degrees of freedom
## Multiple R-squared:  0.2503, Adjusted R-squared:  0.2118 
## F-statistic:  6.51 on 2 and 39 DF,  p-value: 0.003634
```

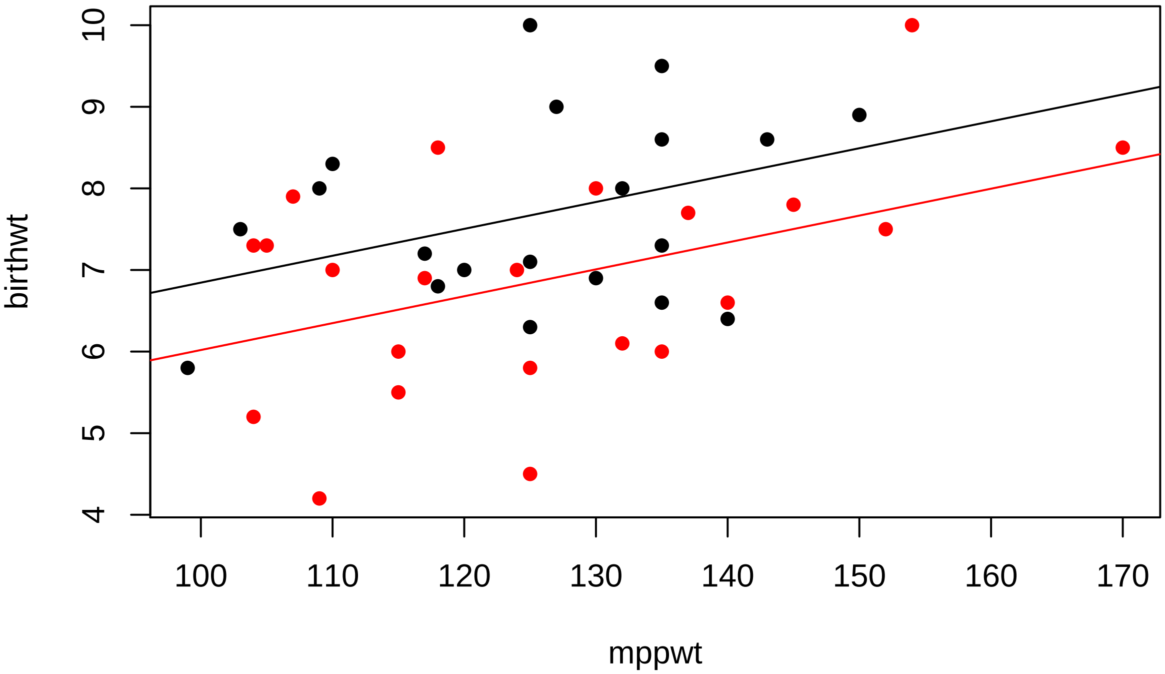
Now all terms are significant at the 5% level. This is the minimal adequate model. The equation for the model is

$$birthwt = 3.549 + 0.0329 \times mppwt - 0.826 \times smoker$$

where *smoker* is a binary variable with values 0 for non-smokers and 1 for smokers. The model has a common slope 0.0329, but different intercepts, 3.549 for non-smokers and 3.549 - 0.826 = 2.723 for smokers. This means that babies from smoking modethers weight on average 0.826 pounds less than babies from non-smoking mothers.

(iii) Draw a scatter plot of `birthwt` against `mppwt` and color the dots according to the value of `smoker`. Add the regression lines for your model. Predict the `birthwt` value for a `mppwt` value of 120 and both values for `smoker`. Add prediction intervals at the 98% level.

```r
plot(birthwt ~ mppwt, col = smoker+1, pch = 16, data = bwt)
c3 <- coef(mod3)
abline(c3[1], c3[2], col = 'black')
abline(c3[1]+c3[3], c3[2], col = 'red')
```



The predicted values with prediction intervals are

```
predict(mod3, data.frame(mppwt = 120, smoker = 0), interval = 'p', level = .98)
```

```
##        fit      lwr      upr
## 1 7.503815 4.565039 10.44259
```

```
predict(mod3, data.frame(mppwt = 120, smoker = 1), interval = 'p', level = .98)
```

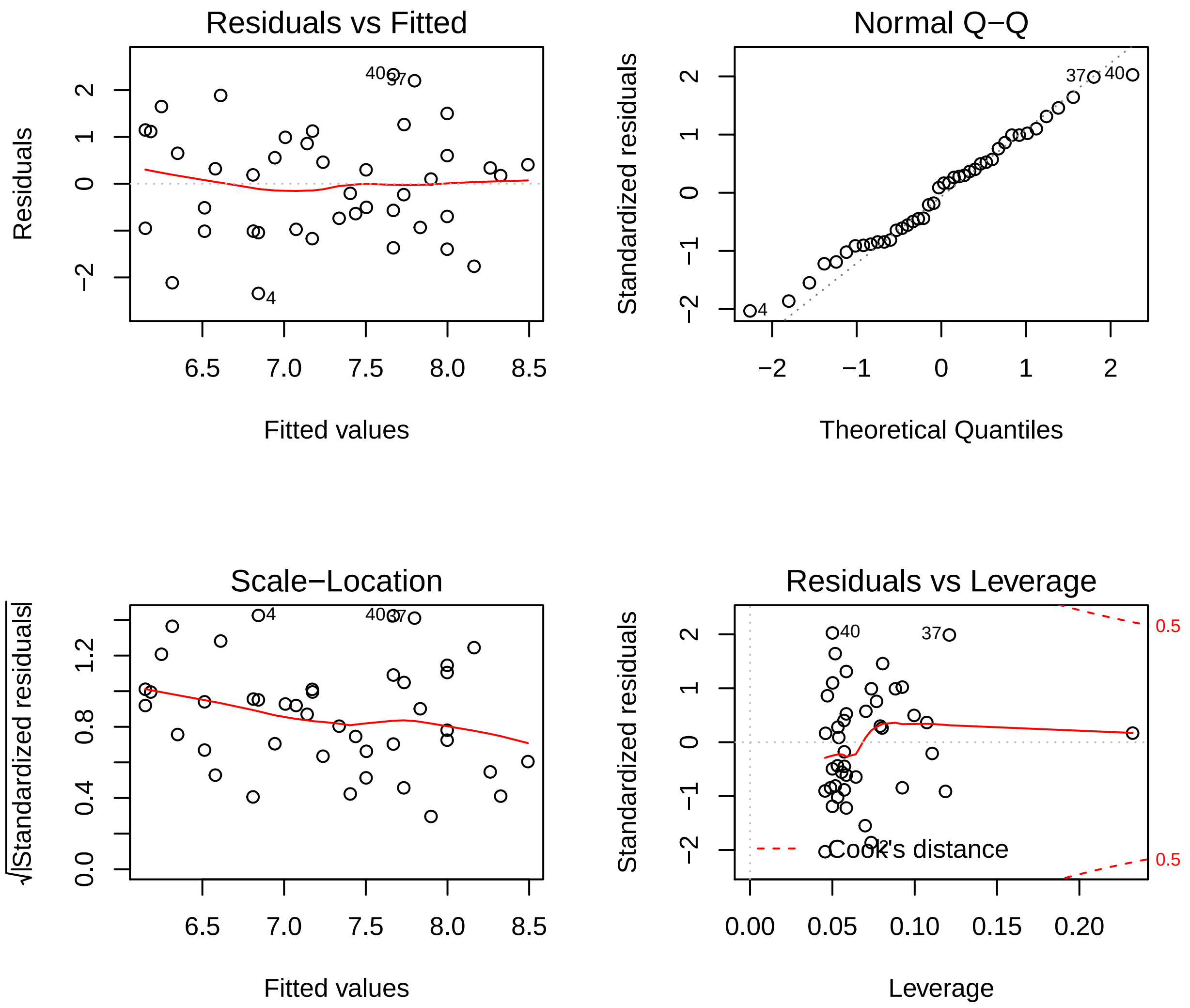
```
##        fit      lwr      upr
## 1 6.678056 3.745004 9.611109
```

(iv) State clearly the assumptions on which the regression model is based. Using graphs and hypothesis tests, do a diagnostic analysis for the model you fitted and verify whether these assumptions are satisfied.

The assumptions for the regression model are that the errors are independent variables with common normal distribution with mean 0 and common variance σ^2. The regressors are assumed to be linearly independent.

We look at the diagnostic plots.

```
par(mfrow=c(2,2))
plot(mod3)
```



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

All assumptions look satisfied except perhaps for homogeneous variances. We use the tests to confirm.

```
shapiro.test(rstandard(mod3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod3)
## W = 0.9816, p-value = 0.7232
```

```
library(car)
ncvTest(mod3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3629979, Df = 1, p = 0.54685
```

The tests confirm that the assumptions for the model are valid.