TMA4267 Linear Statistical Models
Compulsory Exercise Part 2: Linear regression

Due date: 3 March 2017 at 6pm

Two students can hand in one solution together. One student uploads the (handwritten) solutions in Blackboard with the name of both students on the front page, the other only submits "I have handed in my solution together with NN" in Blackboard.

For Problems **a)** and **b)** you may answer all exercise questions based on the supplied information (as you will at the written exam). For the very last part of **c)** you need to use R to fit the reduced model that you choose.

Problem

In a medical study the aim was to explain the etiology of diabetes progression. Data was collected from $n = 442$ diabetes patients, and from each patient the following measurements are available:

- `age` (in years) at baseline
- `sex` (0=female and 1=male) at baseline
- body mass index (`bmi`) at baseline
- mean arterial blood pressure (`map`) at baseline
- six blood serum measurements: total cholesterol (`tc`), ldl cholesterol (`ldl`), hdl cholesterol (`hdl`), `tch`, `ltg`, glucose `glu`, all at baseline,
- a quantitative measurement of disease progression one year after baseline (`prog`)

All measurements except `sex` are continuous.

A multiple linear regression model is fitted to the data set with `prog` as response and all the other measurements as covariates. We call this the *full model.*

a) Refer to the print-out from `summary(full)` in Figure 1 and answer *briefly* the following questions:

(i) For each column (`Estimate, Std.Error, t value, Pr(>|t|)`), write down the underlying mathematical formula that the numerical values are based on.

(ii) How do you interpret the regression coefficient estimate for the intercept? (That is, which values of the covariates would give this as the predicted response?)

(iii) What type of coding is done for the categorical variable `sex`? What does this imply for the interpretation of the model?

(iv) How would you explain to someone unfamiliar with linear regression how the estimated regression coefficient for `bmi` can be interpreted?

(v) Where (in the print-out) can you find the estimated error variance?

(vi) Which of the covariates are found to be *significant* at level 0.05? Write down the null- and alternative hypotheses associated with one such test. What are the assumptions needed for the p-value to be valid?

b) How would you, based on Figures 1 and 2 evaluate the fit of the full model?

Is the regression significant? Write down the null- and alternative hypotheses for this test.

Explain what the number called `Multiple R-squared` in Figure 1 means.

The researchers also want to use the data to fit a prediction model, and want to consider reduced versions of the full model based on best subset model selection.

c) Why might a reduced model have better performance than a full model when the aim is prediction?
Explain briefly what is done in the best subset model selection, and give the reasoning behind the R^2_{adj} and BIC criteria.

Results from using the R^2_{adj} and the BIC criteria are presented in Figures 3 and 4. Based on these results choose a reduced regression model, fit this reduced model in R, and write down the fitted regression model for the model you choose.

Compare the estimated regression parameters and the estimated standard deviations for the full model (Figure 1) and the reduced model that you choose. Explain what you observe.

```
>ds=read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv",
sep=",")
>apply(ds,2,summary)
            age    sex   bmi    map    tc    ldl   hdl  tch   ltg    glu  prog
Min.    19.00 0.0000 18.00  62.00  97.0  41.60 22.00 2.00 1.410  58.00  25.0
1st Qu. 38.25 0.0000 23.20  84.00 164.2  96.05 40.25 3.00 1.860  83.25  87.0
Median  50.00 0.0000 25.70  93.00 186.0 113.00 48.00 4.00 2.005  91.00 140.5
Mean    48.52 0.4683 26.38  94.65 189.1 115.40 49.79 4.07 2.016  91.26 152.1
3rd Qu. 59.00 1.0000 29.28 105.00 209.8 134.50 57.75 5.00 2.170  98.00 211.5
Max.    79.00 1.0000 42.20 133.00 301.0 242.40 99.00 9.09 2.650 124.00 346.0
>pairs(ds,pch=".")
>full=lm(prog~.,data=ds)
>summary(full)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -356.64395   67.01983  -5.321 1.66e-07 ***
age           -0.03529    0.21705  -0.163 0.870910
sex          -22.79233    5.83657  -3.905 0.000109 ***
bmi            5.59548    0.71746   7.799 4.75e-14 ***
map            1.11589    0.22526   4.954 1.05e-06 ***
tc            -1.08286    0.57294  -1.890 0.059428 .
ldl            0.73914    0.53032   1.394 0.164108
hdl            0.36783    0.78274   0.470 0.638648
tch            6.54048    5.95956   1.097 0.273045
ltg          157.17606   36.04811   4.360 1.63e-05 ***
glu            0.28148    0.27332   1.030 0.303661
---
Residual standard error: 54.16 on 431 degrees of freedom
Multiple R-squared:  0.5176,        Adjusted R-squared:  0.5065
F-statistic: 46.25 on 10 and 431 DF,  p-value: < 2.2e-16
> plot(full$fitted,rstudent(full),pch=20)
> qqnorm(rstudent(full),pch=20)
> qqline(rstudent(full),col=2)
> library(nortest)
> ad.test(rstudent(full))
        Anderson-Darling normality test
data:  rstudent(full)
A = 0.37292, p-value = 0.4176
```

Figure 1: R code and print-out for fitting the full model, for **a)** and **b)**.

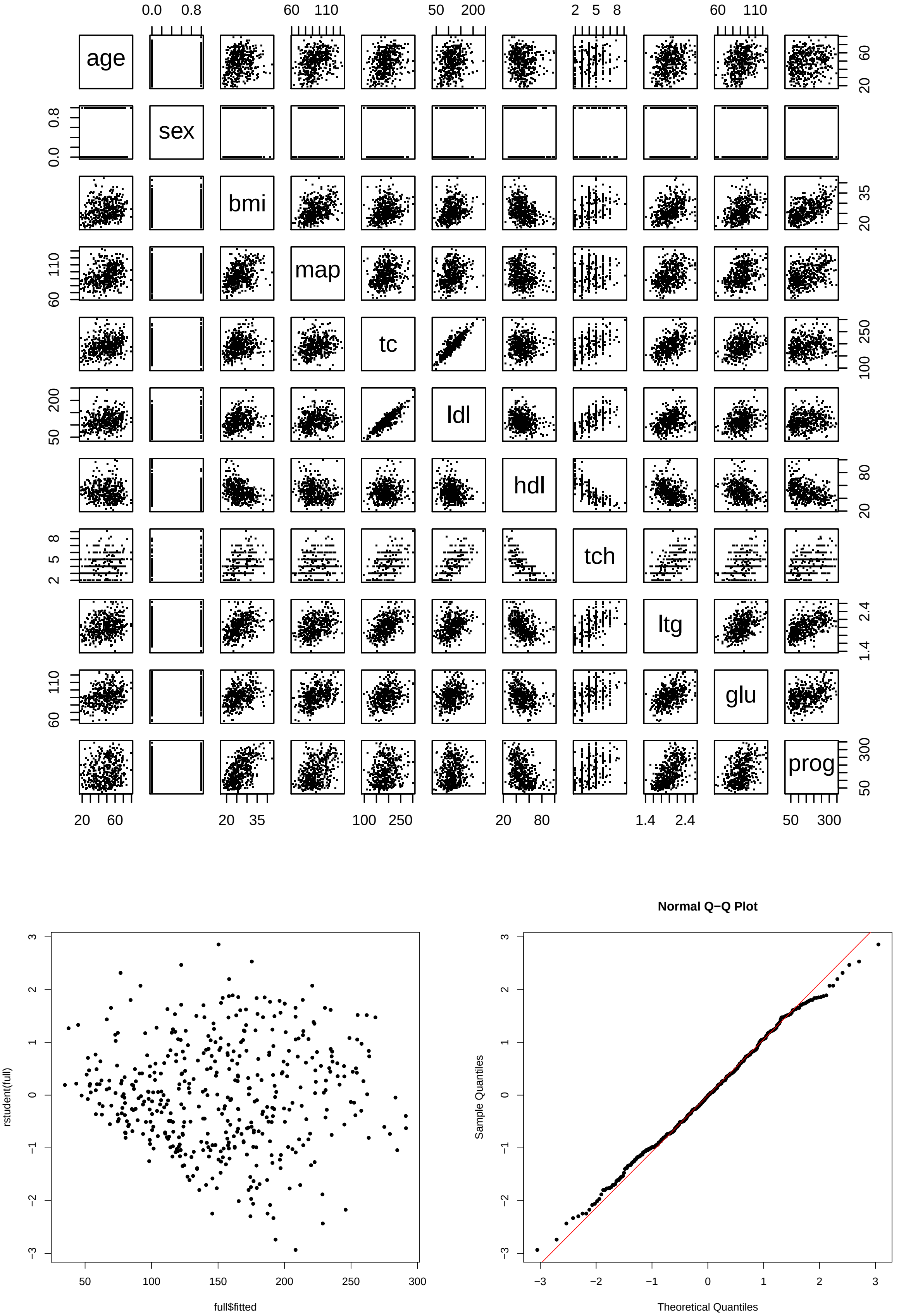

Figure 2: Top row: Scatter plots of all variables in the data set. Bottom row: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for the full model.

```
> library(leaps)
> allsubs = regsubsets(prog~. , data = ds , nvmax = 10)
> allsummary = summary(allsubs)
> allsummary
Subset selection object
Call: regsubsets.formula(prog ~ ., data = ds, nvmax = 10)
10 Variables  (and intercept)
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          age sex bmi map tc  ldl hdl tch ltg glu
1  ( 1 )  " " " " "*" " " " " " " " " " " " " " "
2  ( 1 )  " " " " "*" " " " " " " " " " " "*" " "
3  ( 1 )  " " " " "*" "*" " " " " " " " " "*" " "
4  ( 1 )  " " " " "*" "*" "*" " " " " " " "*" " "
5  ( 1 )  " " "*" "*" "*" " " " " "*" " " "*" " "
6  ( 1 )  " " "*" "*" "*" "*" "*" " " " " "*" " "
7  ( 1 )  " " "*" "*" "*" "*" "*" " " "*" "*" " "
8  ( 1 )  " " "*" "*" "*" "*" "*" " " "*" "*" "*"
9  ( 1 )  " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
> plot(allsummary$bic, xlab="Number of Variables", ylab="BIC", type="l")
> round(allsummary$bic,1)
 [1] -174.1 -253.7 -264.8 -268.9 -277.5 -277.0 -272.2 -267.2 -261.3 -255.2
> which.min(allsummary$bic)
[1] 5
> plot(allsubs,scale="bic",col=gray(seq(0.5,0.95,length=20)))
> plot(allsummary$adjr2, xlab="Number of Variables", ylab="R2adj", type="l")
> round(allsummary$adjr2,4)
 [1] 0.3424 0.4571 0.4766 0.4874 0.5029 0.5081 0.5084 0.5085 0.5076 0.5065
> which.max(allsummary$adjr2)
[1] 8
```

Figure 3: R code and print-out for finding a reduced model in **c)**.

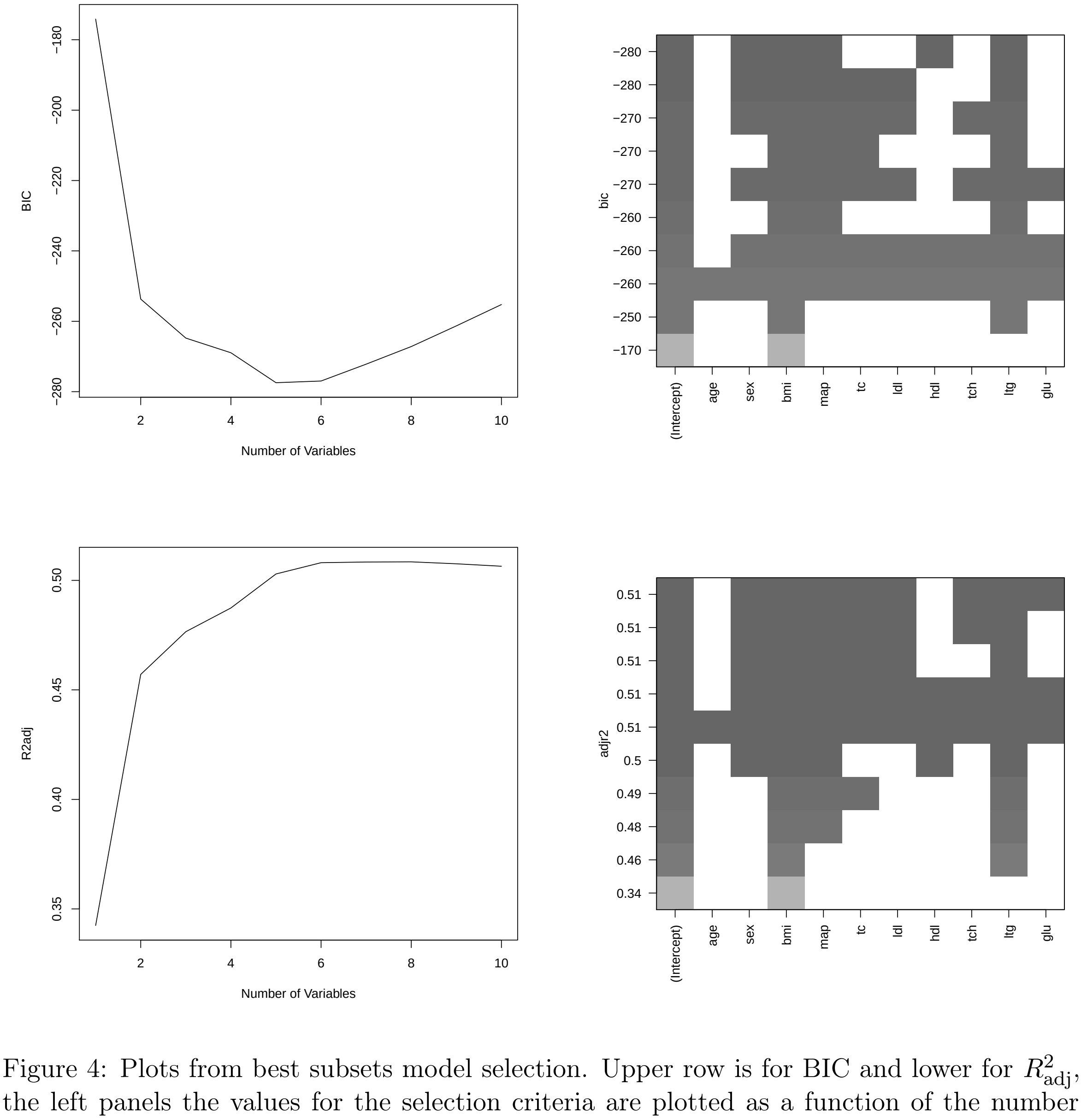

Figure 4: Plots from best subsets model selection. Upper row is for BIC and lower for R^2_{adj}, in the left panels the values for the selection criteria are plotted as a function of the number of covariates in the chosen models and the right panels summarize chosen covariates sorted by the model criteria (white areas=not in model, grey areas=in model).