Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models 2020 spring
Compulsory exercises 2

Problem 1 Diabetes progression

For Problems (a) and (b), you may answer all 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, and you need R for (d) .

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 were available:

`age` (in years) at baseline (start of study)

`sex` (0=female and 1=male) at baseline

`bmi` body mass index at baseline

`map` mean arterial blood pressure at baseline

`tc`, `ldl`, `hdl`, `tch`, `ltg`, `glu` six blood serum measurements: total cholesterol, ldl cholesterol, hdl cholesterol, . . . , glucose, all at baseline

`prog` a quantitative measurement of disease progression one year after baseline

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 *briefly* answer the following questions:

1. For each column `Estimate`, `Std. Error`, `t value`, `Pr(>|t|)`, write down the formula that the numerical values are based on, and explain all quantities used (e.g., what is $\boldsymbol{Y}$?).
2. How do you interpret the estimate for the intercept? (That is, which values of the covariates would give this as the predicted response?)

3. How would you explain to someone unfamiliar with linear regression how the estimated regression coefficient for `bmi` can be interpreted?
4. Where (in the print-out) can you find the estimated error variance? What is the formula for the estimated error variance?
5. 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. In particular, explain how the 10 models presented in Figure 3 was found.

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.

One possible reduced model in (c) is the model including the five covariates `sex`, `bmi`, `map`, `hdl`, `ltg` and an intercept. This means that the five covariates `age`, `tc`, `ldl`, `tch`, `glu` from the full model is not included in the reduced model. We want to investigate whether hypothesis testing in the full model would confirm that the reduced model is preferable.

d) Perform the test

$$H_0 : \beta_{\text{age}} = \beta_{\text{tc}} = \beta_{\text{ldl}} = \beta_{\text{tch}} = \beta_{\text{glu}} = 0 \qquad \text{versus} \qquad H_1 : \text{at least one} \neq 0$$

in the full model. Report a p-value of the test. Comment on the result. Would you prefer the full or the reduced model?

```
> 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.    19.0000 0.0000000 18.00000  62.00000  97.0000  41.6000 22.00000 2.000000 1.410000  58.00000  25.0000
1st Qu. 38.2500 0.0000000 23.20000  84.00000 164.2500  96.0500 40.25000 3.000000 1.860000  83.25000  87.0000
Median  50.0000 0.0000000 25.70000  93.00000 186.0000 113.0000 48.00000 4.000000 2.005000  91.00000 140.5000
Mean    48.5181 0.4683258 26.37579  94.64661 189.1403 115.4391 49.78846 4.070249 2.015747  91.26018 152.1335
3rd Qu. 59.0000 1.0000000 29.27500 105.00000 209.7500 134.5000 57.75000 5.000000 2.170000  98.00000 211.5000
Max.    79.0000 1.0000000 42.20000 133.00000 301.0000 242.4000 99.00000 9.090000 2.650000 124.00000 346.0000
> pairs(ds, pch = ".")
> full <- lm(prog ~ ., data = ds)
> summary(full)

Call:
lm(formula = prog ~ ., data = ds)

Residuals:
     Min       1Q   Median       3Q      Max
-156.308  -38.402   -0.727   38.003  151.606

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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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).

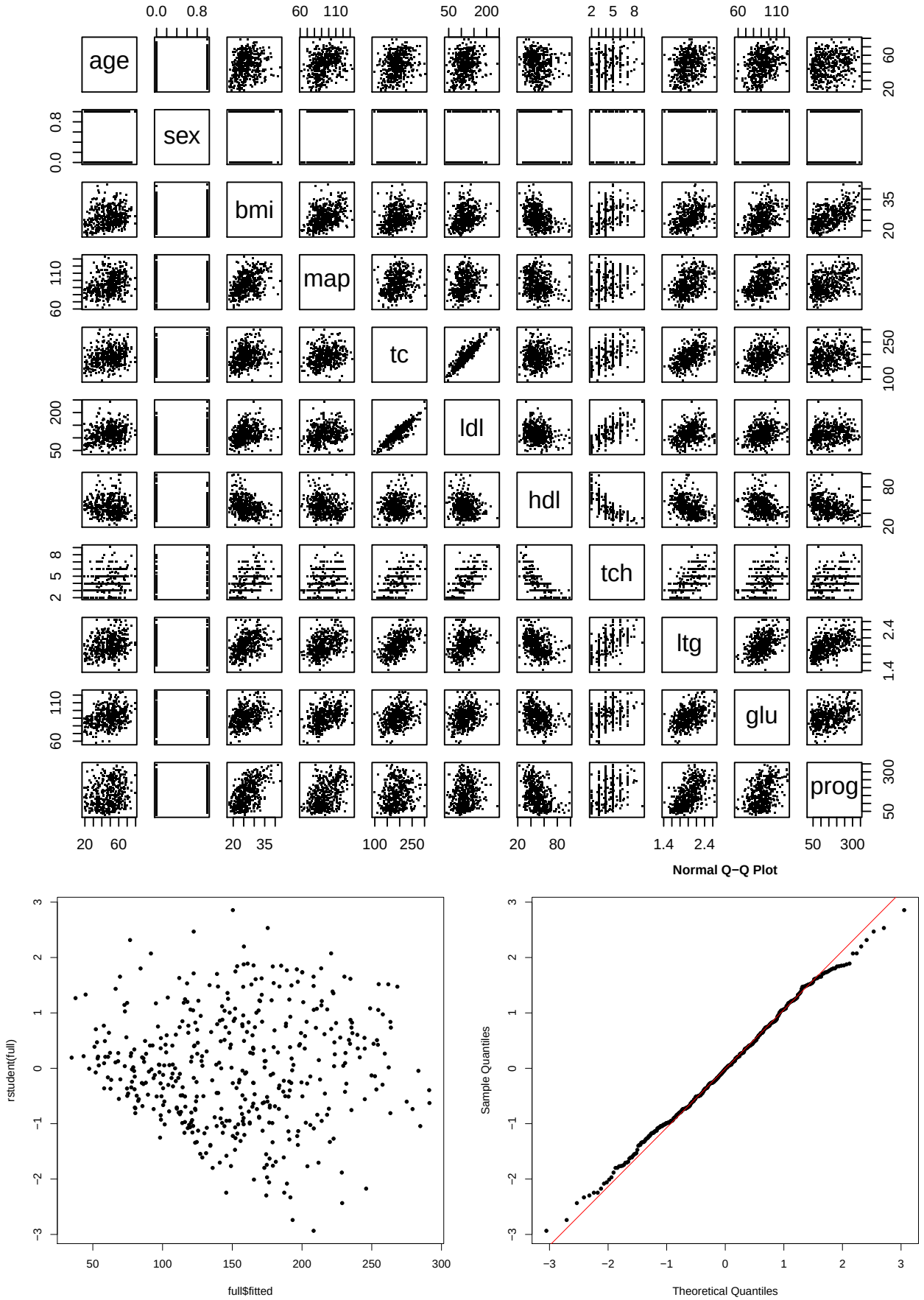

Figure 2: Scatter plots of all variables in the data set (top). Studentized residuals versus fitted values (bottom left). Normal Q–Q plot based on studentized residuals (bottom right).

```
> library(leaps)
> allsubs <- regsubsets(prog ~ . , data = ds , nvmax = 10)
> allsummary <- summary(allsubs)
> allsummary$outmat
          age sex bmi map tc  ldl hdl tch ltg glu
1  ( 1 )  " " " " " "*" " " " " " " " " " " " " "
2  ( 1 )  " " " " " "*" " " " " " " " " " " "*" " "
3  ( 1 )  " " " " " "*" "*" " " " " " " " " "*" " "
4  ( 1 )  " " " " " "*" "*" "*" " " " " " " "*" " "
5  ( 1 )  " " "*" "*" "*" " " " " "*" " " "*" " "
6  ( 1 )  " " "*" "*" "*" "*" "*" " " " " "*" " "
7  ( 1 )  " " "*" "*" "*" "*" "*" " " "*" "*" " "
8  ( 1 )  " " "*" "*" "*" "*" "*" " " "*" "*" "*"
9  ( 1 )  " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
> plot(allsummary$bic,
+      xlab = "Number of Variables",
+      ylab = "BIC",
+      type = "l")
> allsummary$bic
 [1] -174.1108 -253.6898 -264.7696 -268.9392 -277.4512 -276.9716 -272.1793 -267.1786 -261.3082 -255.2440
> which.min(allsummary$bic)
[1] 5
> plot(allsubs, scale = "bic")
> plot(allsummary$adjr2,
+      xlab = "Number of Variables",
+      ylab = "R2adj",
+      type = "l")
> allsummary$adjr2
 [1] 0.3424327 0.4570604 0.4765560 0.4873974 0.5029191 0.5080619 0.5083754 0.5084543 0.5075628 0.5064505
> which.max(allsummary$adjr2)
[1] 8
> plot(allsubs, scale = "adjr2")
>
```

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

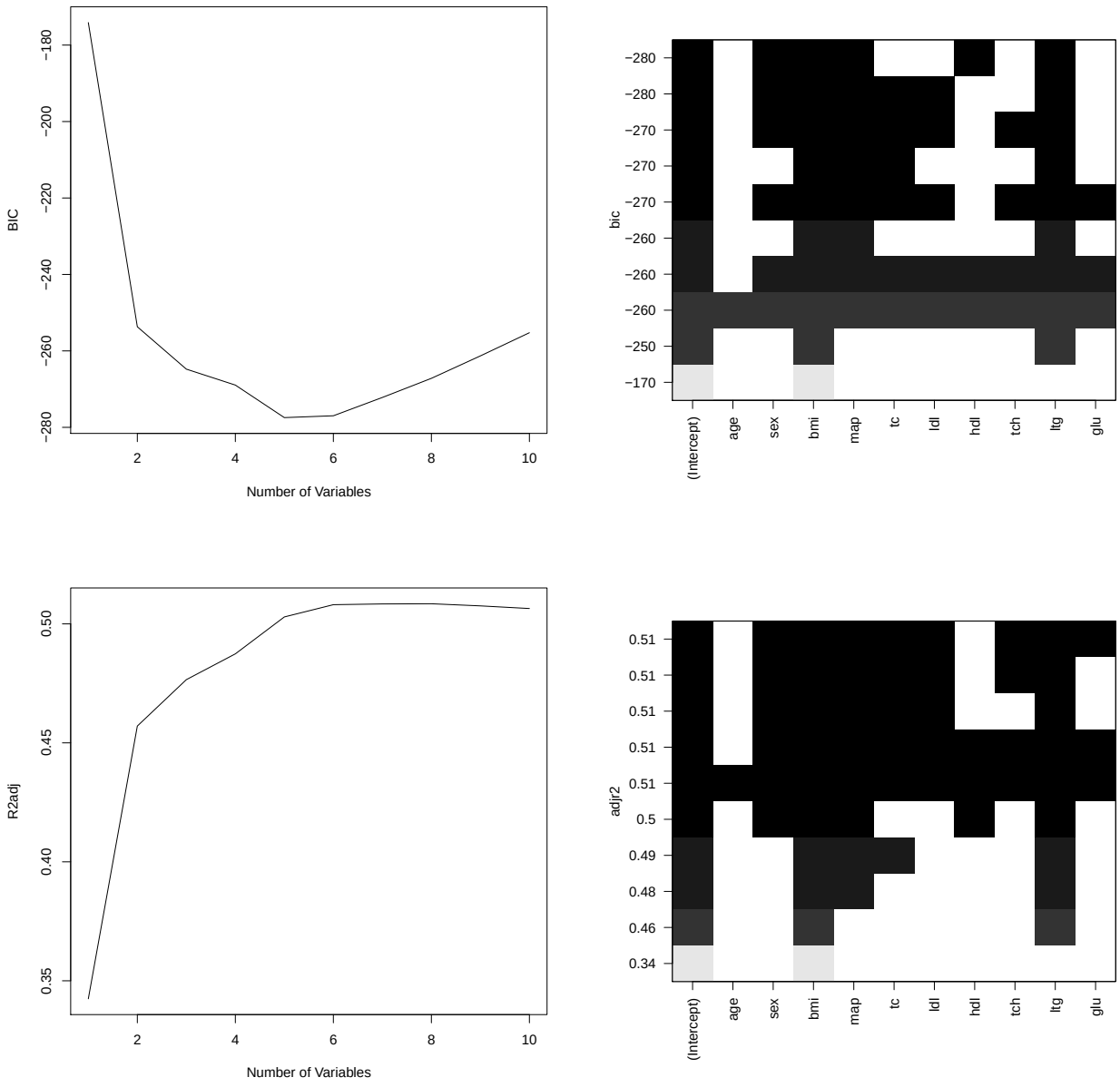

Figure 4: Plots from best subsets model selection. Upper row is for BIC and lower for R^2_{adj}. In the left panels, the values for the selection criteria are plotted as a function of the number of covariates in the chosen models and the right panels summarize chosen covariates sorted by the model criteria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

Problem 2 Multiple testing

In genome-wide association studies, the aim is to test if there is an association between a genetic marker and a trait. We have data from 1000 markers, and for each marker, we perform a hypothesis test,

$$H_0 \colon \beta_j = 0 \qquad \text{versus} \qquad H_1 \colon \beta_j \neq 0,$$

$j = 1, \ldots, 1000$, where $\beta_j = 0$ means that there is no association between the marker and trait (β_j might be a coefficient in a multiple regression model). From hypothesis test j, we calculate a p-value p_j (based on some continuous test statistic). P-values are available to read into R as

```
pvalues <-
  scan("https://www.math.ntnu.no/emner/TMA4267/2018v/pvalues.txt")
```

a) Assume that we reject all null-hypotheses with corresponding p-values below 0.05. How many null-hypotheses do we reject? What is a false positive finding (type I error)? Do we know the number of false positive findings in our data?

b) What is the definition of the familywise error rate, FWER?

What does it mean to control the FWER at level 0.05?

What cut-off on p-values (significance level) should we use if we want to control the FWER at level 0.05 for our data with the Bonferroni method? How many null-hypotheses will we reject with this new cut-off?

c) To see the effect of choosing different cut-offs on p-value on the number of false positive findings we need to know which null hypotheses are true and which are false. Let us assume that the first 900 null hypotheses are true and the last 100 are false. What does this imply about the number of type I and type II errors in (a) and (b)?