Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination: Øyvind Bakke

Phone: 73 59 81 26, 990 41 673

Examination date: 25 May 2018

Examination time (from–to): 9:00–13:00

Permitted examination support material: Yellow stamped A5 sheet with your own handwritten notes, specific basic calculator, *Tabeller og formler i statistikk* (Tapir forlag), *Matematisk formelsamling* (K. Rottmann)

Other information:
In the grading, each of the ten points counts equally. All answers must be justified, and relevant calculations provided.

Language: English

Number of pages: 4

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Date Signature

Problem 1

Assume that $\boldsymbol{X} = (X_1\ X_2)^{\mathrm{T}}$ has a bivariate normal distribution with mean vector $(0\ 0)^{\mathrm{T}}$ and covariance matrix $\left(\begin{smallmatrix} 1 & -0.8 \\ -0.8 & 2 \end{smallmatrix}\right)$.

a) Find the conditional distribution of X_2 given $X_1 = x$.

Assume that $\boldsymbol{Y}$ has a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and non-singular (invertible) covariance matrix Σ.

b) Give a definition of $\Sigma^{1/2}$ such that $(\Sigma^{1/2})^2 = \Sigma$ and $\Sigma^{1/2}$ is non-singular. What is the distribution of $(\Sigma^{1/2})^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$?

c) Derive the distribution of $(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$.

Problem 2

Calory intake was studied for twenty male diabetics. A multiple linear regression model was fitted, with the percentages of total calories obtained from complex carbohydrates as the response, and age, weight and percentage of calories obtained from protein as covariates (data from Dobson and Barnett, *An introduction to generalized linear models, Third edition*). R input and output and some plots are shown in Figure 1.

a) Comment briefly on the model fit. Calculate the total sum of squares (SST), the regression sum of squares (SSR, also called explained sum of squares) and the error sum of squares (SSE, also called residual sum of squares).

b) Explain what best subset selection is. What is the philosophy of model choice criteria, such as Mallows' C_P, and why are the coefficient of determination (R^2) or the error sum of squares (SSE) not suitable as model choice criteria? Which model would you prefer for the carbohydrate data set?

```
> fit<-lm(carbo~age+weight+protein)
> summary(fit)

Call:
lm(formula = carbo ~ age + weight + protein)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
age         -0.11368    0.10933  -1.040  0.31389
weight      -0.22802    0.08329  -2.738  0.01460 *
protein      1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.956 on 16 degrees of freedom
Multiple R-squared:  0.4805,Adjusted R-squared:  0.3831
F-statistic: 4.934 on 3 and 16 DF,  p-value: 0.01297

> rres<-rstudent(fit)
> plot(fit$fitted.values,rres)
> qqnorm(rres)
> qqline(rres)
> library(leaps)
> carbdata<-as.data.frame(cbind(carbo,age,weight,protein))
> best<-regsubsets(carbo~.,data=carbdata)
> summary(best)$which
  (Intercept)   age weight protein
1        TRUE FALSE  FALSE    TRUE
2        TRUE FALSE   TRUE    TRUE
3        TRUE  TRUE   TRUE    TRUE
> summary(best)$cp
[1] 8.201698 3.081179 4.000000
> plot(best,scale="Cp",col=gray(c(0,.2,.4)))
```

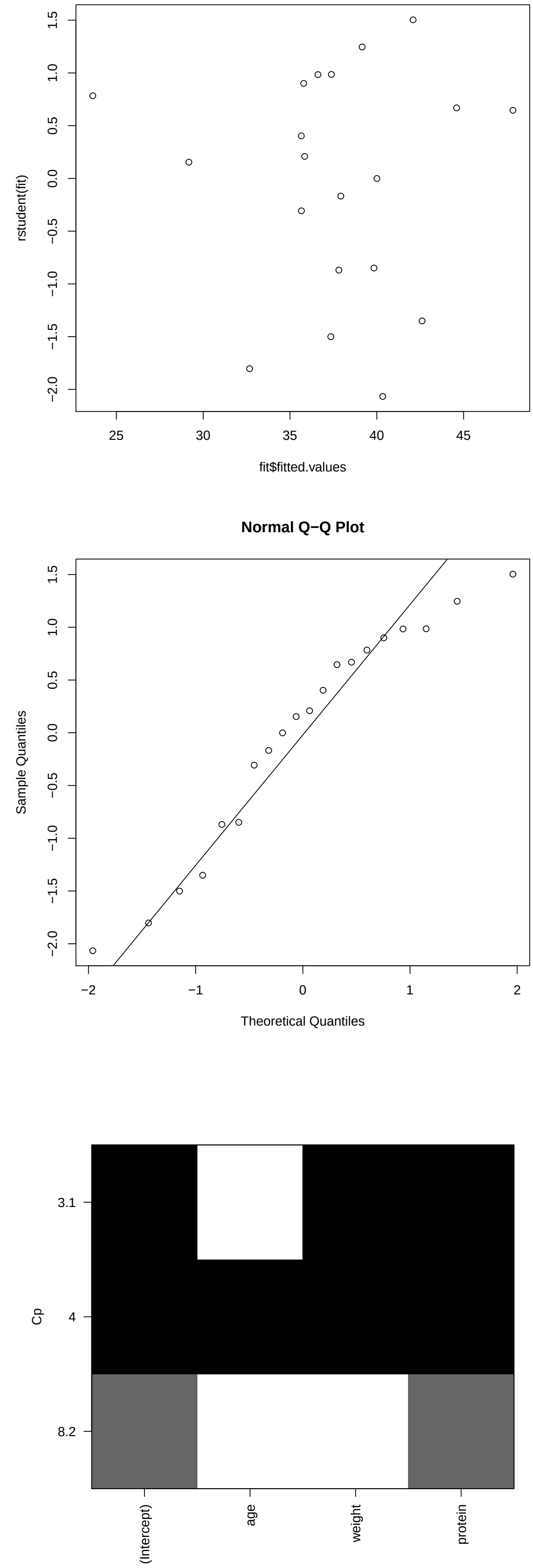

Figure 1: Model from Problem 2a: R input and output (left), residual plot (upper right), normal Q–Q plot (middle right), and a graphical table of best subsets using Mallows' C_P as the statistic for ordering models (lower right). Note that the information of the graphical table is also included in the R output.

Problem 3

A response variable Y_{ij} was measured, using 15 repetitions for each of four levels of a factor. A regression model of the form $Y_{kj} = \beta_j + \epsilon_{kj}$ was assumed, where $k = 1, 2, \ldots, 15$, $j = 1, 2, 3, 4$, and the ϵ_{kj} were independent $N(0, \sigma^2)$.

Another way to formulate the model is $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$, $i = 1, 2, \ldots, 60$, with $x_{ij} = 1$ if the factor was at level j in experiment i and $x_{ij} = 0$ otherwise.

a) Assuming that the factor was at level 1 for $i = 1, \ldots, 15$, at level 2 for $i = 16, \ldots, 30$, at level 3 for $i = 31, \ldots, 45$, and at level 4 for $i = 46, \ldots, 60$, explain how the design matrix X looks (including its dimensions). Show that $(X^{\mathrm{T}}X)^{-1} = \frac{1}{15}I$, with I a 4×4 identity matrix.

The least-squares estimates of β_3 and of β_4 were 1.0902858 and 0.1752633, respectively, and the error sum of squares was SSE = 43.04524.

b) Perform a test in which the null hypothesis is $H_0\colon \beta_3 = \beta_4$ and the alternative hypothesis is $H_1\colon \beta_3 \neq \beta_4$. Use significance level 0.05. You should calculate a test statistic and use its distribution under H_0 to arrive at your conclusion.

A corresponding test was performed for all pairs of coefficients. The p-values are given in the following table.

H_0	$\beta_1 = \beta_2$	$\beta_1 = \beta_3$	$\beta_1 = \beta_4$	$\beta_2 = \beta_3$	$\beta_2 = \beta_4$	$\beta_3 = \beta_4$
p-value	0.0251	0.3698	0.0557	0.0022	0.7297	0.0060

c) What is *family-wise error rate* (FWER)? Suggest a method that keeps the familywise error rate below 0.05 when performing the tests. Which null hypotheses are rejected?

Problem 4

Assume that $\boldsymbol{Y}$ is a random vector with $E\boldsymbol{Y} = X\boldsymbol{\beta}$ and $\operatorname{Cov}\boldsymbol{Y} = \sigma^2 I$, with X a design (model) matrix, $\boldsymbol{\beta}$ a vector of coefficients and I an identity matrix. Let $\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{Y}$ be the least-squares estimator of $\boldsymbol{\beta}$.

Consider another linear estimator, i.e., $\tilde{\boldsymbol{\beta}} = B\boldsymbol{Y}$, where B is a matrix, that is also an unbiased estimator of $\boldsymbol{\beta}$, i.e., $E\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.

a) Find the covariance matrices of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ in terms of σ^2, X and B. Show that $\boldsymbol{\beta} = BX\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, so that $BX = I_p$, an identity matrix.

Let $M = \sigma\Big(B - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Big)$.

b) Show that $MM^{\mathrm{T}} = \operatorname{Cov}\tilde{\boldsymbol{\beta}} - \operatorname{Cov}\hat{\boldsymbol{\beta}}$. What can you conclude about the variance in each component of $\tilde{\boldsymbol{\beta}}$ compared to the variance of the corresponding component of $\hat{\boldsymbol{\beta}}$?