NTNU
Noregs teknisk-naturvitskaplege universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk
Institutt for matematiske fag

English
Contact person:
John Tyssedal 73593534/41645376

Exam in TMA4267 Linear Statistical Models
Saturday June 5. 2010
Time 09.00-13.00

Permitted aids: A yellow stamped A-5 sheet with your own handwritten notes.
Tabeller og formler i statistikk (Tapir forlag). K. Rottman: Matematisk formelsamling.
Calculator HP30S or Citizen SR-270X.

Problem 1

Let $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \begin{bmatrix} 4 \\ -3 \\ 1 \end{bmatrix}$ and. $\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & -3/2 \\ 0 & -3/2 & 5 \end{bmatrix}$

a) Find the distribution of $X_1 + X_2 + X_3$ and of X_2 given $X_1 = x_1$ and $X_3 = x_3$

Help (For $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$, we have

$$\left(\boldsymbol{X}_1 \middle| \boldsymbol{X}_2 = \boldsymbol{x}_2 \right) \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \right)$$

b) Find the eigenvalues and the eigenvectors of $\boldsymbol{\Sigma}$.

Determine a 3×3 matrix $\boldsymbol{A}$ such that for $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \boldsymbol{AX}$, Y_1, Y_2 and Y_3 are independent.

Problem 2

A study was conducted to determine whether lifestyle change could replace medication in reducing blood pressure among hypertensives (people with very high blood pressure). The factors considered were

H = Healthy diet with an exercise program
M = Medication
N = No intervention.

12 people with very high blood pressure participated in the experiment. Four of these (randomly selected) were put on a healthy diet and followed an exercise program, four were given medication and the rest got no treatment. The response, Y, is the change in blood pressure after some period. The data is given in the table below

H	M	N
-32	-11	-6
-21	-19	5
-26	-23	-11
-16	-5	14

Assume that a model for the response is given by:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i=1,2,3,\ j=1,2,3,4 \text{ where } \varepsilon_{ij} \sim N\left(0,\sigma^2\right),\ i=1,2,3,\ j=1,2,3,4 \text{ and}$$

independent.

Below is given some commands and output from the statistical computer software R.

```
> y=c(-32,-21,-26,-16,-11,-19,-23,-5,-6,5,-11,14)
> treat=c("H","H","H","H","M","M","M","M","N","N","N","N")
> lmtreat=lm(y~treat-1)
> summary(lmtreat)

Call:
lm(formula = y ~ treat - 1)

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
treatH   -23.75       4.45  -5.338  0.00047 ***
treatM   -14.50       4.45  -3.259  0.00986 **
treatN     0.50       4.45   0.112  0.91300

> mean(y)
[1] -12.58333
```

```
> summary(aov(y~treat))
            Df  Sum Sq Mean Sq F value  Pr(>F)
treat        2 1198.17  599.08  7.5647 0.01182 *
Residuals    9  712.75   79.19

-> treatdat=data.frame(y,treat)
> TukeyHSD(aov(y~treat, treatdat))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ treat, data = treatdat)

$treat
     diff        lwr      upr     p adj
M-H  9.25 -8.319065 26.81906 0.3489386
N-H 24.25  6.680935 41.81906 0.0097708
N-M 15.00 -2.569065 32.56906 0.0941763
--
```

a) Use this output to do the following:

Test $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$ against H_1: at least one $\alpha_i, i = 1,2,3$ is different from 0.

Use a 5% level of significance.

Calculate estimates for α_1, α_2 and α_3.

Is there any form of treatment (diet and exercise and/or medication) that significantly differs from no intervention? Explain your answer.

The pretreatment body mass index (BMI= $\frac{weight}{(height)^2}$) was also calculated. In order to find out if BMI had any influence it was decided to perform a regression analysis. Therefore three regression variables $x1$, $x2$ and $x3$, were constructed, and the response values are now labeled as $y_i, i = 1,2,\cdots,12$. The response values and regression variables are given below:

i:	1	2	3	4	5	6	7	8	9	10	11	12
y_i:	-32	-21	-26	-16	-11	-19	-23	-5	-6	5	-11	14
$x1_i$:	0	0	0	0	1	1	1	1	0	0	0	0
$x2_i$:	0	0	0	0	0	0	0	0	1	1	1	1
$x3_i$ (BMI):	27.3	22.1	26.1	27.8	19.2	26.1	28.6	23	28.1	25.3	26.7	22.3

Some output from a regression analysis performed with R is given below:

```
> lm2=lm(y ~x1+x2+x3)
> summary(lm2)
Call:
lm(formula = y ~ x1 + x2 + x3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.4999    20.5491   1.095  0.30541
x1            6.3846     5.3386   1.196  0.26597
x2           23.8470     5.1926   4.593  0.00177 **
x3           -1.7909     0.7829  -2.287  0.05147 .

Residual standard error: 7.339 on 8 degrees of freedom
Multiple R-squared: 0.7745,     Adjusted R-squared: 0.6899
F-statistic: 9.159 on 3 and 8 DF,  p-value: 0.005759
```

b)

What null-hypothesis is tested with the F-statistics given above?
What will be the conclusion using a 5% level of significance?
Use the output to calculate $\sum_{i=1}^{12}(y_i - \hat{y}_i)^2$, $\sum_{i=1}^{12}(\hat{y}_i - \overline{y}_i)^2$ and $\sum_{i=1}^{12}(y_i - \overline{y})^2$.

An estimated model without the variable $x1$, gave the following output from R.

```
> lm2=lm(y~x2+x3)
> summary(lm2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.190     19.675   1.585  0.14738
x2            20.781      4.622   4.496  0.00150 **
x3            -2.011      0.779  -2.581  0.02965 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.513 on 9 degrees of freedom
Multiple R-squared: 0.7342,     Adjusted R-squared: 0.6751
F-statistic: 12.43 on 2 and 9 DF,  p-value: 0.002574
```

c) Perform both a t-test and a F-test to test whether $x1$ can explain a significant portion of the variation in the response given that $x2$ and $x3$ are in the model. Use a 5% level of significance.
Does this model provide any additional insight into the treatment of high blood pressure compared to the analysis performed in 2a)? Explain your answer.

Problem 3

Let Y be a response variable, x a regression variable and ε a random variable of errors. α and β are coefficients.

a) Decide for each of the models i), ii) and iii) if they are linear regression models or not and eventually what transformations is needed in order to have a linear regression model. Explain your answer.

i) $Y = x^{\beta} \cdot \varepsilon$

ii) $Y = \alpha + \beta\sqrt{x} + \varepsilon$

iii) $Y = \dfrac{x}{\alpha + (\beta + \varepsilon)x}$

Now consider the linear regression model written in matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y}$ is a $n \times 1$ vector of random variables, $\boldsymbol{X}$ a $n \times (k+1)$ matrix and $\boldsymbol{\beta}$ a $(k+1) \times 1$ vector of parameters. Also assume $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$. Further let $\boldsymbol{J} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ and $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t$.

b) Show that $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent matrices.
Let $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the least square estimator of $\boldsymbol{\beta}$. Show that $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ are independent.

c) The sum of squares for residuals $SS_E = \boldsymbol{Y}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. Explain why $\dfrac{SS_E}{\sigma^2}$ is $\chi^2(n\text{-}k\text{-}1)$.

Under the assumption $\beta_1 = \beta_2 = \cdots \beta_k = 0$ it can be shown that $\dfrac{SS_R}{\sigma^2} = \dfrac{\boldsymbol{Y}^t\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}}{\sigma^2}$ also is χ^2-fordelt. Argue why $\dfrac{SS_R/k}{SS_E/(n\text{-}k\text{-}1)}$ than has a F-distribution with k and $n\text{-}k-1$ degrees of freedom. (Hint. You can use that $\boldsymbol{H1} = \boldsymbol{1}$ if $\boldsymbol{1}$ is a vector of 1's).