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

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 joint density of X_1 and X_2.

Find the distribution of the vector given by: $\begin{bmatrix} X_1 \\ X_2 - 3X_3 \\ 3X_2 + X_3 \end{bmatrix}$.

Problem 2

In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. Researchers at the University of Bergen conducted a series of experiments to study the factors that influence the voltage (Y) required to separate water from the oil. Seven independent variables (regression variables/factors), each set to two values were investigated. Data from 16 conducted experiments are given in the table on the next page.

Disperse Phase Volume x_1 (%)	Salinity x_2 (%)	Temperature x_3 $\left(^\circ\mathrm{C}\right)$	Time delay x_4 (hours)	Surfactant concentration x_5 (%)	Span: Triton x_6	Solid Particles x_7 (%)	Voltage Y (kw/cm)
40	1	4	0.25	2	0.25	0.5	0.64
80	1	4	0.25	4	0.25	2	0.80
40	4	4	0.25	4	0.75	0.5	3.20
80	4	4	0.25	2	0.75	2	0.48
40	1	23	0.25	4	0.75	2	1.72
80	1	23	0.25	2	0.75	0.5	0.32
40	4	23	0.25	2	0.25	2	0.64
80	4	23	0.25	4	0.25	0.5	0.68
40	1	4	24	2	0.75	2	0.12
80	1	4	24	4	0.75	0.5	0.88
40	4	4	24	4	0.25	2	2.32
80	4	4	24	2	0.25	0.5	0.40
40	1	23	24	4	0.25	0.5	1.04
80	1	23	24	2	0.25	2	0.12
40	4	23	24	2	0.75	0.5	1.28
80	4	23	24	4	0.75	2	0.72

A regression analysis was performed and some output from the computer package R is given below:

```
> lmv=lm(y~x1+x2+x3+x4+x5+x6+x7)
> summary(lmv)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.591491   0.747886   0.791  0.45182
x1          -0.020500   0.006650  -3.083  0.01506 *
x2           0.170000   0.088671   1.917  0.09151 .
x3          -0.015263   0.014001  -1.090  0.30738
x4          -0.008421   0.011201  -0.752  0.47368
x5           0.460000   0.133006   3.458  0.00859 **
x6           0.520000   0.532024   0.977  0.35700
x7          -0.126667   0.177341  -0.714  0.49538
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.532 on 8 degrees of freedom
Multiple R-squared:  0.78, Adjusted R-squared: 0.5874
F-statistic: 4.051 on 7 and 8 DF,  p-value: 0.03401
```

a) Use the output from R and answer the following.
Which hypothesis is tested with the F-statistic? What is the conclusion using a 5% level of significance?
Verify the values for R^2 (Multiple R-squared) and adjusted R^2 given in the output.
Which variable would be the first one to be removed in a backward elimination?

The researchers decided to go for a model with the three regression variables x_1, x_2 and x_5. Some output from R is given below:

```
> lmv1=lm(y~x1+x2+x5)
> summary(lmv1)

Call:
lm(formula = y ~ x1 + x2 + x5)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38500    0.60016   0.641  0.53326
x1          -0.02050    0.00643  -3.188  0.00780 **
x2           0.17000    0.08574   1.983  0.07076 .
x5           0.46000    0.12861   3.577  0.00380 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5144 on 12 degrees of freedom
Multiple R-squared: 0.6914,     Adjusted R-squared: 0.6143
F-statistic: 8.963 on 3 and 12 DF,  p-value: 0.002171
```

b) Write down the estimated model. Give an estimate for how an increase in salinity of one percent will affect the voltage.
Write down a test for testing if the variables x_3, x_4, x_6 and x_7 are significant given that the others are in the model. Perform the test. Use a 5% level of significance.

Since each of the regression variables have only two levels these can be transformed to the values -1 and +1. The corresponding design matrix is given on the next page.

x_1 (A)	x_2 (B)	x_3 (C)	x_4 (D)	x_5	x_6	x_7	Y
-1	-1	-1	-1	-1	-1	-1	0.64
1	-1	-1	-1	1	-1	1	0.80
-1	1	-1	-1	1	1	-1	3.20
1	1	-1	-1	-1	1	1	0.48
-1	-1	1	-1	1	1	1	1.72
1	-1	1	-1	-1	1	-1	0.32
-1	1	1	-1	-1	-1	1	0.64
1	1	1	-1	1	-1	-1	0.68
-1	-1	-1	1	-1	1	1	0.12
1	-1	-1	1	1	1	-1	0.88
-1	1	-1	1	1	-1	1	2.32
1	1	-1	1	-1	-1	-1	0.40
-1	-1	1	1	1	-1	-1	1.04
1	-1	1	1	-1	-1	1	0.12
-1	1	1	1	-1	1	-1	1.28
1	1	1	1	1	1	1	0.72

c) The regression variables x_1, x_2, x_3 and x_4 are here renamed to factor A, B, C and D respectively. Explain how x_1 has been transformed. What are the generators for x_5 (E), x_6 (F) and x_7 (G)? What are the defining relations? What is the resolution of this design?

Estimated effects obtained from this design is given below.

```
> effects
(Intercept)          A1          B1          C1          D1          E1
       1.92       -0.82        0.51       -0.29       -0.20        0.92
         F1          G1       A1:B1       A1:C1       A1:D1       A1:E1
       0.26       -0.19       -0.47        0.11        0.16       -0.48
      A1:F1       A1:G1       B1:C1       B1:D1       B1:E1       B1:F1
      -0.16        0.15          NA        0.13          NA          NA
      B1:G1       C1:D1       C1:E1       C1:F1       C1:G1       D1:E1
         NA          NA          NA          NA          NA          NA
      D1:F1       D1:G1       E1:F1       E1:G1       F1:G1
         NA          NA          NA          NA          NA
```

d) Is the high value for the effect of factor A reasonable compared to the one obtained in 2b)? Explain your answer.
Assume the effects with the five largest absolute values are significant. How will the interpretation of how the factors A, B and E affect the response then differ from the interpretation obtained from the model in 2b)?

Problem 3

In a one way analysis of variance with k treatments and n observations for each treatment the model for the response is given by: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i=1,2,\ldots,k$, $j=1,2,\ldots,n$ where $\varepsilon_{ij} \sim N\left(0,\sigma^2\right)$, $i=1,2,\ldots,k$, $j=1,2,\ldots,n$ and independent. Let $\overline{Y}_{..} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}$ and $\overline{Y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$. Define $\boldsymbol{J}_n = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \cdots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n\times n}$ and $\boldsymbol{J}^* = \begin{bmatrix} \frac{1}{n}\boldsymbol{J}_n & 0 & \cdots & 0 \\ 0 & \frac{1}{n}\boldsymbol{J}_n & \cdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{n}\boldsymbol{J}_n \end{bmatrix}_{nk\times nk}$, i.e. a block-diagonal matrix with $\frac{1}{n}\boldsymbol{J}_n$ on the diagonal and all other elements equal to zero.

a) What does it mean that all $\alpha_i, i=1,2,\cdots,k$ are zero.

   Write down an estimator for $\alpha_i, i=1,2,\cdots,k$.

b) Let $\boldsymbol{I}$ be an $nk\times nk$ identity matrix and $\boldsymbol{J}$ a $nk\times nk$ matrix of 1's .

   Show that $\boldsymbol{I}-\boldsymbol{J}^*$ and $\boldsymbol{J}^* - \frac{1}{nk}\boldsymbol{J}$ are idempotent and symmetric.

   Show that $\left(\boldsymbol{I}-\boldsymbol{J}^*\right)\left(\boldsymbol{J}^* - \frac{1}{nk}\boldsymbol{J}\right) = 0$.

c) With $\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n} \\ \vdots \\ Y_{k1} \\ Y_{k2} \\ \vdots \\ Y_{kn} \end{bmatrix}$ the treatment sum of squares can be written as $\boldsymbol{Y}^t\left(\boldsymbol{J}^* - \frac{1}{nk}\boldsymbol{J}\right)\boldsymbol{Y}$ and the residual sum of squares as $\boldsymbol{Y}^t\left(\boldsymbol{I}-\boldsymbol{J}^*\right)\boldsymbol{Y}$.

   Explain how this can be used in constructing a test statistic for testing the hypothesis $H_0:\ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. What is the rejection region for this test?