NTNU
Noregs teknisk-naturvitskaplege universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk
Institutt for matematiske fag

English
Contact during exam:
John Tyssedal 73593534/41645376

Exam in TMA4267 Linear statistical models
Wednesday August 8. 2012
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

Assume $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^t = [1, -1, 2]$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{bmatrix}$.

a) Which pairs of random variables are independent? Find the marginal distribution of X_1. What is the distribution of $Y = X_1 + 3X_2 - 2X_3$?

b) Find the distribution of X_1 given that X_3 is known. Write down a regression model with X_1 as response variable and X_3 as regression variable. What are the parameters in this model?

Hint:

(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

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

Problem 2

A silica company performed an experiment in order to find out the effects of four factors on extracting silica from quartz. The four factors are:

x_1 - quartz size

x_2 - time

x_3 - temperature

x_4 - concentration of sodium hydroxide.

The response, Y, is the amount of extracted silica measured on a logarithmic scale.

a) Assume they decided to have just two levels for each of the four factors and that they made no replicates. How many runs did they need to perform? Which effects are possible to estimate? Assume that third and higher order effects are negligible. Explain how it is then possible to construct a test to find out which effects that are significant.

They believed that some of the factors could have a quadratic effect and they therefore wanted to find out if this could be true. The performed experiment with response values written down in standard form is given below. Notice that this is a two-level experiment augmented with two runs for each of the four factors + four center runs (runs where each factor has a value mid between high and low level). This is an example of a *central composite design*.

x_1	x_2	x_3	x_4	y
-1	-1	-1	-1	4.190
1	-1	-1	-1	5.333
-1	1	-1	-1	4.812
1	1	-1	-1	5.886
-1	-1	1	-1	5.333
1	-1	1	-1	6.225
-1	1	1	-1	5.905
1	1	1	-1	6.574
-1	-1	-1	1	4.533
1	-1	-1	1	5.642
-1	1	-1	1	5.199
1	1	-1	1	6.094
-1	-1	1	1	5.628
1	-1	1	1	6.374
-1	1	1	1	6.127
1	1	1	1	6.702
-2	0	0	0	4.615
2	0	0	0	6.690
0	-2	0	0	5.220
0	2	0	0	6.385
0	0	-2	0	4.796
0	0	2	0	6.693
0	0	0	-2	5.416
0	0	0	2	6.182
0	0	0	0	5.775
0	0	0	0	5.802
0	0	0	0	5.781
0	0	0	0	5.796

A regression analysis performed with R gave the following output for a model with first and second order terms. In this output `x1x2` is the column for the interaction between x_1 and x_2. The same is true for `x1x3` and so on. $x1^2$ is the column with squared x_1 column values. Similarly for $x2^2$, $x3^2$ and $x4^2$.

```
lm(formula = y ~ x1 + x2 + x3 + x4 + x1x2 + x1x3 + x1x4 + x2x3 +
    x2x4 + x3x4 + x1^2 + x2^2 + x3^2 + x4^2, data = kdata)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.788500   0.046095 125.579  < 2e-16 ***
x1            0.468875   0.018818  24.916 2.34e-12 ***
x2            0.265458   0.018818  14.107 2.94e-09 ***
x3            0.457208   0.018818  24.296 3.22e-12 ***
x4            0.148875   0.018818   7.911 2.53e-06 ***
x1x2         -0.042312   0.023047  -1.836  0.08934 .
x1x3         -0.083688   0.023047  -3.631  0.00305 **
x1x4         -0.028312   0.023047  -1.228  0.24105
x2x3         -0.034062   0.023047  -1.478  0.16324
x2x4         -0.009437   0.023047  -0.409  0.68885
x3x4         -0.028312   0.023047  -1.228  0.24105
x1^2         -0.048969   0.018818  -2.602  0.02191 *
x2^2         -0.011469   0.018818  -0.609  0.55273
x3^2         -0.025969   0.018818  -1.380  0.19086
x4^2         -0.012344   0.018818  -0.656  0.52329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09219 on 13 degrees of freedom
Multiple R-squared: 0.9914,     Adjusted R-squared: 0.9822
F-statistic: 107.3 on 14 and 13 DF,  p-value: 4.489e-11
```

b) Is the regression significant? Explain your answer. Calculate the sum of squares for the residuals, SS_E, and the sum of squares for regression, SS_R, from the information given in the output.

c) Write down the estimated model you get by only considering terms that are significant using a 5% level of significance.
Explain why the estimates of first order terms and cross terms (x1x2, x1x3, …) do not depend on which quadratic terms that are in the model. Are the estimates of the effects of the quadratic terms dependent on which terms that are in the model?

The analysis performed above assumes that third and higher order effects are negligible and can be included in the error. Some statisticians that were consulted recommended to perform an additional analysis where the error variance was estimated from the response values for the four center runs only, since these had the same expected value independent of which model were fitted.

With natural definitions the model with an intercept, four first order terms, six cross terms and four quadratic terms can be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is a 28×15 matrix. Let $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{t}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{t}}$.

Let us now especially consider the four center runs. Define a 28×28 matrix $\boldsymbol{J}_c = \begin{bmatrix} \boldsymbol{I}_{24\times 24} & \boldsymbol{0}_{24\times 4} \\ \boldsymbol{0}_{4\times 24} & \frac{1}{4}\boldsymbol{J}_4 \end{bmatrix}$

where $\boldsymbol{I}_{24}$ is a 24×24 identity matrix and $\boldsymbol{J}_4$ a 4×4 matrix with only ones. The residuals $\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{Y}$ can then be written as $\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{Y} = \left(\boldsymbol{I} - \boldsymbol{J}_c\right)\boldsymbol{Y} + \left(\boldsymbol{J}_c - \boldsymbol{H}\right)\boldsymbol{Y}$

d) It can be shown that $\boldsymbol{H}\boldsymbol{J}_c = \boldsymbol{H}$. Use this result to show that the matrices $\left(\boldsymbol{I} - \boldsymbol{J}_c\right)$ and $\left(\boldsymbol{J}_c - \boldsymbol{H}\right)$ are projection matrices. What is the rank of the matrix $\left(\boldsymbol{I} - \boldsymbol{J}_c\right)$?

e) Explain why $\left(\boldsymbol{I} - \boldsymbol{J}_c\right)\boldsymbol{Y} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ Y_{25} - \bar{Y}_c \\ Y_{26} - \bar{Y}_c \\ Y_{27} - \bar{Y}_c \\ Y_{28} - \bar{Y}_c \end{bmatrix}$, where $\bar{Y}_c$ is the average response for the four center runs.

What is the distribution of the quadratic form defined by $\frac{SS_{EC}}{\sigma^2} = \frac{\boldsymbol{Y}^t\left(\boldsymbol{I} - \boldsymbol{J}_c\right)\boldsymbol{Y}}{\sigma^2}$?

Here $\mathrm{Var}(Y) = \sigma^2$. Explain your answer.

f) A possible test statistics for testing whether the regression is significant or not is then given by $\frac{SS_R}{k_1} \Big/ \frac{SS_{EC}}{k_2}$. What is the distribution of this test statistics and what are the values for k_1 og k_2? Explain your answer. What is the conclusion when $SS_{EC} = 0.000477$?