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
Saturday May 25. 2013
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.

Problem1

Assume that the random vector $\boldsymbol{X}=\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}$ is trivariate normal distributed with mean vector $\boldsymbol{\mu}=\begin{bmatrix}2\\6\\4\end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma}=\begin{bmatrix}1&0&1\\0&2&-1\\1&-1&3\end{bmatrix}$.

a) Find out which of the random variables X_1 and X_2 that are most correlated (in absolute value) with X_3. What is the distribution of the vector $\boldsymbol{Z}=\begin{bmatrix}X_2-X_1\\X_3-X_1\end{bmatrix}$?

A company is measuring three quality characteristics in order to control the quality of a product. Their respective random variables can be arranged in a random vector $\boldsymbol{X}=\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}$. Long time experience makes it reasonable to assume that the random vector $\boldsymbol{X}$ is trivariate normal distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ as given on page 1.

The company wants a simple procedure for controlling their products by just having to consider a bivariate vector instead of a trivariate one. The eigenvalues λ_1, λ_2 and λ_3 of $\boldsymbol{\Sigma}$ and their respective eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and $\boldsymbol{e}_3$ are given below as output from R.

```
$values
   [1] 3.8793852 1.6527036 0.4679111

$vectors
              [,1]       [,2]       [,3]
   [1,] -0.2931284 -0.4490988  0.8440296
   [2,]  0.4490988 -0.8440296 -0.2931284
   [3,] -0.8440296 -0.2931284 -0.4490988
```

b) Define the bivariate vector $\boldsymbol{Y} = \begin{bmatrix} Y_1 = \boldsymbol{e}_1^t \boldsymbol{X} \\ Y_2 = \boldsymbol{e}_2^t \boldsymbol{X} \end{bmatrix}$. Why is $\boldsymbol{Y}$ bivariate normal distributed?

Show that Y_1 and Y_2 are independent. How much of the total variation in $\boldsymbol{X}$ can be explained by $\boldsymbol{Y}$?

To test whether their product should pass the control or not the company want to use $D = \left(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right)^t \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right)$ as their test statistic.

c) What is the distribution of D? How will you decide whether the product should pass the control or not? Explain your answer.
To be able to compare this procedure for controlling the product to using the whole trivarate distribution they want to simulate n observations from the trivariate distribution of $\boldsymbol{X}$. To their disposal they only have a random number generator that can produce observations from independent normal distributed variables. Suggest a way to help them with the simulations.

Problem 2

Mount Etna erupted in 1669, 1780 and 1865. When molten lava hardens, small magnetic particles will take the direction of the Earth's magnetic field. Three blocks of lava were examined from each of these eruptions and the declination of the magnetic field in these blocks was measured. The results are given in the table below.

1669 (Y1)	1780 (Y2)	1865 (Y3)
57.8	57.9	52.7
60.2	55.2	53.0
60.3	54.8	49.4

Let $Y_{1j},\ j=1,2,3,\ Y_{2j},\ j=1,2,3$ and $Y_{3j},\ j=1,2,3$ be the declinations in block *j* for the three years 1669, 1780 and 1865 respectively. We assume the observations follow the following model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1,2,3,\ \text{j=1,2,3}$$

where $\sum_{i=1}^{3} \alpha_i = 0$ and $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$ and independent. An output of an analysis with R is given below. The three years 1669, 1780 and 1865 are in the output denoted as Y1, Y2 and Y3.

```
     > summary(aov(y~year))
                 Df Sum Sq Mean Sq F value  Pr(>F)   
     year         2  90.03   45.01   15.28 0.00442 **
     Residuals    6  17.67    2.95                   
     ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       Tukey multiple comparisons of means
         95% family-wise confidence level

     Fit: aov(formula = y ~ year, data = vuldat)

     $year
                diff        lwr         upr     p adj
     Y2-Y1 -3.466667  -7.766308  0.83297418 0.1055367
     Y3-Y1 -7.733333 -12.032974 -3.43369248 0.0035909
     Y3-Y2 -4.266667  -8.566308  0.03297418 0.0514713
```

a) Do these data suggest that the declination of the earth's magnetic field has shifted over the time period spanned by the eruptions?
Can you conclude that there has been a shift within a time period of one hundred year?
Answers the questions by writing down the appropriate hypothesis and perform the necessary comparisons. You can use significance level $\alpha = 0.05$ in your evaluation of significance.

Another way to analyse these data is as follows. Define

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{bmatrix}.$$

We then have a multiple linear regression model of the form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. An output from a regression analysis performed by R is on the next page.

```
        >lm(formula = y ~ x1 + x2 + x3 - 1)

        Coefficients:
           Estimate Std. Error t value Pr(>|t|)
        x1  59.4333     0.9909   59.98 1.44e-09 ***
        x2  55.9667     0.9909   56.48 2.07e-09 ***
        x3  51.7000     0.9909   52.18 3.33e-09 ***

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

        Residual standard error: 1.716 on 6 degrees of freedom
        Multiple R-squared: 0.9994,     Adjusted R-squared: 0.9991
        F-statistic:  3170 on 3 and 6 DF, p-value: 5.482e-10
```

b) How will you interpret β_1, β_2 and β_3 in terms of μ, α_1, α_2 and α_3? Are the least square estimators $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ independent ? Explain your answer.

This way of analysis also suggest a way to test significance and we want to investigate this further:

Define $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{t}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{t}$. Define also

$\boldsymbol{J}_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{J}^{*} = \begin{bmatrix} \frac{1}{3}\boldsymbol{J}_3 & 0 & 0 \\ 0 & \frac{1}{3}\boldsymbol{J}_3 & 0 \\ 0 & 0 & \frac{1}{3}\boldsymbol{J}_3 \end{bmatrix}$. Finally let $\boldsymbol{I}$ be a $\boldsymbol{I}$ 9×9 identity matrix and let $\boldsymbol{J}$ be a 9×9 matrix with only ones. Three possible ways to partition the data are given below.

1: $\left(\boldsymbol{I} - \frac{1}{9}\boldsymbol{J}\right)\boldsymbol{Y} = \left(\boldsymbol{I} - \boldsymbol{J}^{*}\right)\boldsymbol{Y} + \left(\boldsymbol{J}^{*} - \frac{1}{9}\boldsymbol{J}\right)\boldsymbol{Y}$

2: $\left(\boldsymbol{I} - \frac{1}{9}\boldsymbol{J}\right)\boldsymbol{Y} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{Y} + \left(\boldsymbol{H} - \frac{1}{9}\boldsymbol{J}\right)\boldsymbol{Y}$

3: $\boldsymbol{Y} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{Y} + \boldsymbol{H}\boldsymbol{Y}$

c) Find the matrix $\boldsymbol{H}$. Explain why expressions 1 and 2 are identical. What is the rank of the respective matrices in the three expressions (you can use known results from the theory)? Explain your answers.

d) To quadratic forms are given by $\boldsymbol{Y}^t\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{Y}$ and $\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)^t\boldsymbol{H}\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)$. Are these independent? What is the distribution of $\dfrac{\boldsymbol{Y}^t\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{Y}}{\sigma^2}$? What is the distribution of $\dfrac{\boldsymbol{Y}^t\boldsymbol{HY}}{\sigma^2}$ assuming $\beta_1=\beta_2=\beta_3=0$? Explain your answers.

e) Let $\mu_i=\mu+\alpha_i,\ i$=1,2,3. Suppose you want to test the following two null hypothesis against their alternatives

$$H_0^1:\mu_1=\mu_2=\mu_3$$

$$H_0^2:\mu_1=\mu_2=\mu_3=0$$

Write down a test statistics for both situations and give the conclusions.

The standard way to define R^2 is $R^2=\dfrac{\boldsymbol{Y}^t\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}}{\boldsymbol{Y}^t\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}}$, where n is the number of responses.

What is R^2 for these data with this definition?