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
Tuesday May 22. 2011 Correct year is 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

A person plans to invest some money in a fund. He assumes that the value of the fund he will invest in at time t, X_t, is normally distributed with mean μ_t and variance $k\mu_t^2$ where k is a given constant. To simplify notation let X_1 be the value of the fund at time t_1 and let X_2 be the value of the fund at time t_2, $t_2 > t_1$. We write $X_1 \sim N\left(\mu_1, k\mu_1^2\right)$ and $X_2 \sim N\left(\mu_2, k\mu_2^2\right)$. Let the correlation between X_1 and X_2 be $\rho > 0$.

a) Assume that X_1 and X_2 are binormal distributed . Find the covariance between X_1 and X_2 and write down the covariance matrix for the stochastic vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Find the distribution of the stochastic vector $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 - X_1 \end{bmatrix}$.

b) $Y_2 = X_2 - X_1$ represents the return of the fund in the time interval $\left[t_1, t_2\right]$. Find the distribution of Y_2 given that $Y_1 = X_1$ is known. Will you recommend the person to invest in the fund when X_1 is above or when X_1 is below its expected value. Explain your answer.

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

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

Since the variance is proportional to the square of the expected value the person was recommended to perform the analysis with transformed variables. Let $Z_1 = \ln\left(X_1\right)$ and $Z_2 = \ln\left(X_2\right)$.

c) Find approximated variances of Z_1 and Z_2. Show that the correlation between Z_1 and Z_2 is approximately ρ and find an approximation to the covariance matrix for the stochastic vector $\begin{bmatrix} Z_2 - Z_1 \\ Z_1 \end{bmatrix}$.

Hint: (If $Z = g\left(X\right)$ you can use that $Z \approx g\left(\mu_X\right) + g'\left(\mu_X\right)\left(X - \mu_X\right)$)

Problem2

A statistician is considering two models that can be fitted to a set of data. For the first model, model 1, she assumes a polynomial model for the response, i.e.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2} + \cdots + \beta_r x_i^{\,r} + \varepsilon_i,\ i{=}1,2,\cdots,n \qquad \text{Model 1}$$

where $\varepsilon_i \sim N\left(0, \sigma^2\right),\ i{=}1,2,\cdots,n$ and independent. With natural definitions this model can be written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{X}$ being a $n \times p$ matrix, $p = r + 1$.

As the alternative model, model 2, she assumes, since x in this case takes k distinct values and she has m observations for each x-value, that the same response follows the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i{=}1,2,\ldots,k,\ j{=}1,2,\ldots,m \qquad \text{Model 2}$$

where $\sum_{i=1}^{k} \alpha_i = 0$ and $\varepsilon_{ij} \sim N\left(0, \sigma^2\right),\ i{=}1,2,\ldots,k,\ j{=}1,2,\ldots,m$ and independent. Note that $n = km$.

Now let $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ for model 1 and $\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1m} \\ \vdots \\ Y_{k1} \\ Y_{k2} \\ \vdots \\ Y_{km} \end{bmatrix}$ for model 2. The two vectors of random variables are written such that they are identical. Also define $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{t}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{t}}$,

$$\boldsymbol{J}_m = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \cdots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{m\times m} \quad \text{and } \boldsymbol{J}^* = \begin{bmatrix} \frac{1}{m}\boldsymbol{J}_m & 0 & \cdots & 0 \\ 0 & \frac{1}{m}\boldsymbol{J}_m & \cdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{m}\boldsymbol{J}_m \end{bmatrix}_{mk\times mk}$$

Further Let $\boldsymbol{I}$ be an $mk\times mk$ identity matrix and $\boldsymbol{J}$ a $mk\times mk$ matrix of 1's .
Two ways of partitioning the variables are:

For model 1: $\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = \left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{Y} + \left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}$

For model 2: $\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = \left(\boldsymbol{I}-\boldsymbol{J}^*\right)\boldsymbol{Y} + \left(\boldsymbol{J}^*-\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}$

a) Show that the matrices $\left(\boldsymbol{I}-\boldsymbol{H}\right)$ and $\left(\boldsymbol{I}-\boldsymbol{J}^*\right)$ are idempotent and symmetric.

What is the rank of the matrices $\left(\boldsymbol{I}-\boldsymbol{H}\right)$, $\left(\boldsymbol{I}-\boldsymbol{J}^*\right)$, $\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)$ and $\left(\boldsymbol{J}^*-\frac{1}{n}\boldsymbol{J}\right)$.

There are now two ways of investigating if the mean of Y is dependent on x. One assuming that model 1 is the correct one and one assuming that model 2 is the correct one.

b) Write down H_0 and the alternative hypothesis H_1 for both models.
Which test statistics will be used in each case.

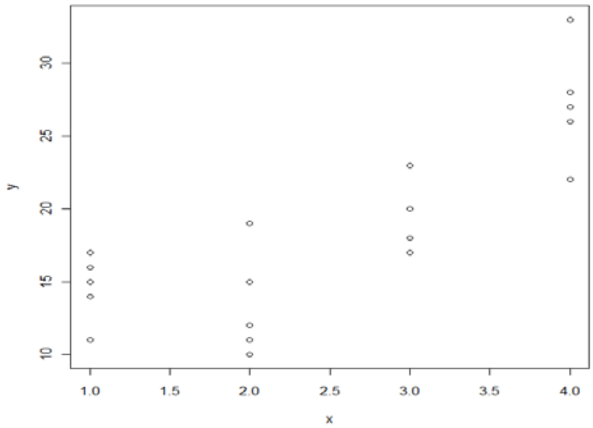
c) Derive the distribution for the test statistics for model 2 using known theoretical results.

Problem 3

A company wishes to test four different package designs for a new breakfast cereal. Twenty stores, with approximately equal sales volumes, were selected for the investigation. Each store was randomly assigned one of the package designs such that each type of package design was given to five stores. Other relevant factors such as price, amount and location of shelf space and advertising, were kept the same for all the stores. After a certain given time period, sales in number of cases were registered in each of the twenty stores. The data are given in the Table below.

Package design			
1	2	3	4
11	12	23	27
17	20	20	33
16	15	18	22
14	19	17	26
15	11	20	28

A plot of number of cases sold by package design is given below.



As a first modeling in order to find out how the number of cases sold depend on package design a regression analysis was performed with a similar model as model 1 in Problem 2 where x = type of package design and x2= x^2 were used as regression variables. x takes the values 1,2,3,4. Output from R is given below.

```
lm(formula = y ~ x + x2)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.7000     3.9068   4.787 0.000172 ***
x            -6.6000     3.5641  -1.852 0.081502 .
x2            2.2000     0.7017   3.135 0.006030 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.138 on 17 degrees of freedom
Multiple R-squared: 0.7763,     Adjusted R-squared: 0.7499
F-statistic: 29.49 on 2 and 17 DF,  p-value: 2.97e-06
```

a) Is the regression significant on a 1% level? Explain your answer. Show that the residual sum of squares, SS_E, equals 167.4 (or approximately 167.4) for this model. What is the regression sum of squares? Use the model with both regression variables to estimate expected number of cases sold for package design 4.

One alternative is to perform a one-way analysis of variance (similar to model 2 in Problem 2). Output from R follows below:

```
lm(formula = y ~ Package design- 1)

Coefficients:
                  Estimate Std. Error t value  Pr(>|t|)
Package design 1   14.600      1.407  10.376  1.64e-08 ***
Package design 2   13.400      1.407   9.523  5.40e-08 ***
Package design 3   19.600      1.407  13.929  2.31e-10 ***
Package design 4   27.200      1.407  19.330  1.62e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.146 on 16 degrees of freedom
```

```
Multiple R-squared: 0.9795,     Adjusted R-squared: 0.9744

F-statistic: 191.5 on 4 and 16 DF,  p-value: 2.713e-13


> summary(aov(y~Package design))

                    Df  Sum Sq Mean Sq F value    Pr(>F)

Package design       3   589.8   196.6  19.859   1.214e-05 ***

Residuals           16   158.4     9.9

Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Will you conclude that number of cases sold is dependent on package design from this analysis? Explain your answer.
The company wishes to establish a prediction interval for number of cases sold with package design 4 in a time period of same length as in the investigation.
Explain why $\overline{Y}_{4.} = \frac{1}{5}\sum_{j=1}^{5} Y_{4j}$ is a natural predictor for the number of cases sold with package design 4 using a similar model as model 2 in Problem 2. Which of the two models give the shortest $(1-\alpha)100\%$ prediction interval?

Hint: The 20 diagonal elements of the matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{t}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{t}$ are in given order:

```
0.19 0.19 0.19 0.19 0.19 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11
0.11 0.11 0.19 0.19 0.19 0.19 0.19
```