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
Friday May 20 2011
Time 15.00-19.00

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

An exam consists of two tasks. For each task a score is given and the final mark is decided from the sum of the scores for each task. Let X_1 be the score on task 1 and let X_2 be the score on task 2. For each student we shall assume that $X_1 \sim N\left(\mu,\sigma^2\right)$, that $X_2 \sim N\left(k\,\mu,k^2\,\sigma^2\right)$ and that $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is binormal distributed. Let the correlation between X_1 and X_2 be ρ.

a) What is the distribution of the random vector $\boldsymbol{Y} = \begin{bmatrix} X_1 + X_2 \\ X_1 \end{bmatrix}$? Explain your answer. Find the expectation of $\boldsymbol{Y}$, $\boldsymbol{\mu_Y}$, and the covariance matrix of $\boldsymbol{Y}$, $\Sigma_{\boldsymbol{Y}}$.

b) Find $E\left[X_1 + X_2 \middle| X_1 = x_1\right]$. What is the expected final score for a student that has a score of one point more than what is expected on the first task when $k = 4$ and $\rho = 1/2$? Find the first principal component of $\Sigma_{\boldsymbol{X}} = Cov\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ when $k = 1$ and $\rho = 1/2$.
How much of the variance of $X_1 + X_2$ is explained by the first principal component?

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

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

Problem 2

A treatment plant is interested in how ph-value and three different polymers affect the amount of solid material in the filter. The three polymers are categorical variables and given by the indicator variables z_1 and z_2. The coding for these are as follows

Indicators	z_1	z_2
Polymer 1	1	0
Polymer 2	0	1
Polymer 3	0	0

The collected data are given in the table below..

solid material	ph	z_1	z_2
292	6.5	1	0
329	6.9	1	0
352	7.8	1	0
378	8.4	1	0
392	8.8	1	0
410	9.2	1	0
198	6.7	0	1
227	6.9	0	1
277	7.5	0	1
297	7.9	0	1
364	8.7	0	1
375	9.2	0	1
167	6.5	0	0
225	7.0	0	0
247	7.2	0	0
268	7.6	0	0
288	8.7	0	0
342	9.2	0	0

It was first tried out to fit a model with solid material as response and with ph, z_1 and the cross product term between ph and z_1 as regression variables. Output from R with this model, model 1, is given below

```
> model 1=lm(solid.mat~ph+z1+ph*z1, data=material)
> summary(model 1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)

(Intercept) -215.410     47.623  -4.523 0.000477 ***
ph            62.942      6.095  10.327 6.26e-08 ***
z1           254.827     81.087   3.143 0.007196 **
ph:z1        -22.680     10.225  -2.218 0.043605 *

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

Residual standard error: 19.63 on 14 degrees of freedom
Multiple R-squared: 0.9368, Adjusted R-squared: 0.9232
F-statistic: 69.12 on 3 and 14 DF,  p-value: 1.236e-08
```

a) Is the regression significant? Explain your answer. What is the interpretation of "Multiple R-squared"? Show that the sum of squares for the residuals, SS_E, is 5394.72 (or approximately 5394.72). What is the sum of squares for regression?

It was then decided to augment the model with z_2 and the cross product between ph and z_2. For this model, model 2, we get the following output from R.

```
> model 2=lm(solid.mat~ph+z1+ph*z1+z2+ph*z2, data=material)
> summary(model 2)
Call:
lm(formula = avsatt.mat ~ ph + z1 + ph * z1 + z2 + ph * z2, data = material)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -158.275     48.517  -3.262   0.0068 **
ph            53.824      6.253   8.607 1.76e-06 ***
z1           197.692     68.795   2.874   0.0140 *
z2          -108.740     71.051  -1.530   0.1518
ph:z1        -13.561      8.737  -1.552   0.1466
ph:z2         17.394      9.090   1.914   0.0798 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.59 on 12 degrees of freedom
Multiple R-squared: 0.9701,     Adjusted R-squared: 0.9576
F-statistic: 77.76 on 5 and 12 DF,  p-value: 1.016e-08
```

b) What is the sum of squares for the residuals with model 2? Perform a test in order to find out if the terms that contain z_2 i.e. , z_2 and ph:z2, can be taken out. Use a 5% level of significance.

In model 2, z_2 , is the least significant variable. By removing this we get the following model, model 3.

```
> model 3=lm(solid.mat~ph+z1+phz1+phz2)
> summary(model 3)
Call:
lm(formula = avsatt.mat ~ ph + z1 + ph*z1 + ph*z2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -208.978     37.229  -5.613 8.43e-05 ***
ph            60.309      4.830  12.487 1.30e-08 ***
z1           248.395     63.328   3.922  0.00175 **
ph:z1        -20.047      8.025  -2.498  0.02669 *
ph:z2          3.581      1.134   3.159  0.00755 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.32 on 13 degrees of freedom
Multiple R-squared: 0.9642,     Adjusted R-squared: 0.9532
F-statistic: 87.57 on 4 and 13 DF,  p-value: 2.886e-09
```

c) Perform and construct a test in order to investigate if the cross product term between ph and z_2 is significant given that ph, z_1 and the cross product term between ph and z_1 also are in the model. Use a 5 % level of significance. Use model 3 to write down a fitted model for each of the three polymers.

Problem 3

In matrix form the linear regression model can be written as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y}$ is a $n \times 1$ vector of stochastic variables, $\boldsymbol{X}$ is a $n \times (k+1)$ matrix and $\boldsymbol{\beta}$ a $(k+1) \times 1$ vector of parameters. Assume also that $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$. The least square estimator for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t \boldsymbol{Y}$. Let $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t$

a) Show that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t \boldsymbol{\varepsilon}$ and that $Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1}$. Let $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the vector of residuals. Show that $\boldsymbol{e} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}$.

For a quadratic form $\boldsymbol{Y}^t \boldsymbol{A} \boldsymbol{Y}$, where the matrix $\boldsymbol{A}$ is symmetric we have that $E\left(\boldsymbol{Y}^t \boldsymbol{A} \boldsymbol{Y}\right) = tr\left(\boldsymbol{AV}\right) + \boldsymbol{\mu}^t \boldsymbol{A} \boldsymbol{\mu}$ where $\boldsymbol{\mu} = E\left(\boldsymbol{Y}\right)$ and $\boldsymbol{V} = Cov\left(\boldsymbol{Y}\right)$.

b) Explain why the sum of squares for the residuals, SS_E, can be written as $\boldsymbol{\varepsilon}^t \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}$ and use the result above to show that $E\left(SS_E\right) = \left(n\text{-}k-1\right)\sigma^2$. Suggest an unbiased estimator for σ^2.

We shall now assume that $\boldsymbol{\varepsilon} \sim \boldsymbol{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$

c) Show that $\frac{SS_E}{\sigma^2} \sim \chi^2\left(n-k-1\right)$ and that $\hat{\boldsymbol{\beta}}$ and SS_E are independent.