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
August 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.

Problem 1

Assume that the random vector $\boldsymbol{V} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$ is trivariate normal distributed with mean vector $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$.

a) Find the joint distribution of $U = X + Y$ and $V = Y + Z$. Also determine for which values of a and b, U and $W = aY + bZ$ are independent.

Problem 2

An experiment was conducted in order to investigate how the addition of sand and carbon both measured in % affected the hardness and strength of casting. We will now only consider the hardness. The results from the experiment performed is given below. Notice that both factors have three levels.

Sand\Carbon	0%	0.25%	0.5%
0%	61, 63	69, 69	67, 69
15%	67, 69	69, 74	69, 74
30%	65, 74	74, 72	74, 74

Let Y_{ijk} denote the hardness obtained with the *i*-th level of sand, the *j*-th level of carbon in the *k*-th replication, We will assume that the following model is appropriate for the data.

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i = 1,2,3$, $j = 1,2,3$, $k = 1,2$ where all ε_{ijk} are independent and $N(0,\sigma^2)$. In addition $\sum_{i=1}^{3}\alpha_i = \sum_{j=1}^{3}\beta_j = \sum_{i=1}^{3}(\alpha\beta)_{ij} = \sum_{j=1}^{3}(\alpha\beta)_{ij} = 0$

a) What kind of experiment has been performed? Explain what the parameters in the model mean. Write also down the three hypothesis that normally are of particular interest when data from such a model is analysed.

An output from an analysis with R is given below.

```
> lmcasting=lm(Hardness~Sand*Carbon,casting)
> anova(lmcasting)
Analysis of Variance Table

Response: Hardness
            Df  Sum Sq Mean Sq F value  Pr(>F)
Sand         2 106.778  53.389  6.5374 0.01764 *
Carbon       2  87.111  43.556  5.3333 0.02967 *
Sand:Carbon  4   8.889   2.222  0.2721 0.88870
Residuals    9  73.500   8.167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Write down the test statistics that are used in this output and give the conclusion on the three hypothesis of particular interest (suggested in a). Use a 5% level of significance. In which order should the hypothesis tests be performed.

The three levels of Sand are denoted S1, S2 and S3 in increasing order. Similarly the three levels of Carbon are denoted C1, C2 and C3. Below is some output from R. First we find the average hardness of all observations

```
> mean(Hardness)
  69.61111
```

Then we have calculated the average hardness for each level of sand

```
> Sandmean=lm(Hardness~Sand -1)
> summary(Sandmean)
```

```
Coefficients:

         Estimate Std. Error t value Pr(>|t|)
SandS1   66.333      1.372   48.34   <2e-16 ***
SandS2   70.333      1.372   51.25   <2e-16 ***
SandS3   72.167      1.372   52.59   <2e-16 ***
```

Thereafter we have calculated the average hardness for each level of Carbon

```
> Carbonmean=lm(Hardness~Carbon -1)
> summary(Carbonmean)

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
CarbonC1    66.50       1.45   45.87   <2e-16 ***
CarbonC2    71.17       1.45   49.09   <2e-16 ***
CarbonC3    71.17       1.45   49.09   <2e-16 ***
---
```

Finally we find the average of hardness for each level combination of Sand and Carbon

```
> tapply(Hardness,list(Sand,Carbon),mean)

     C1   C2   C3
S1 62.0 69.0 68.0
S2 68.0 71.5 71.5
S3 69.5 73.0 74.0
```

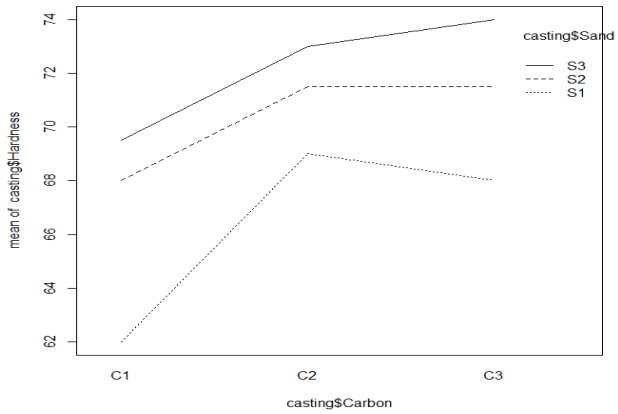
c) Find estimates for α_1, β_2 and $(\alpha\beta)_{33}$. Perform a test to investigate if the hardness on the level combination (S3, C3) is larger than on the level combination (S1, C1). Use a 5% level of significance.

We observe that the distance between S2 and S1 is the same as between S3 and S2. We also observe that the distance between C2 and C1 is the same as between C3 and C2. Hence it is possible to define new factor columns by transforming S1 to -1, S2 to 0 and S3 to 1 and similarly for C1, C2 and C3. Let us denote the new factor columns obtained for Sand and Carbon as x_1 and x_2. Their transposed values are given below:

x_1^t = (-1, -1, 0, 0, 1, 1, -1, -1, 0, 0, 1, 1, -1, -1, 0, 0, 1, 1)

x_2^t = (-1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)

From the interaction plot given below we also observe that for each level of sand there seems to be some curvature in hardness when carbon increases.



Let us define $x_{22} = x_2^2$. Two regression analysis was performed where Hardness was regressed on x_1 and x_2 and x_1, x_2 and x_{22} respectively. The output from R is given below

```
> model_1=lm(Hardness~x1+x2+x1**2)
> summary(model_1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.6111     0.6350  109.63  < 2e-16 ***
x1            2.9167     0.7777    3.75  0.00193 **
x2            2.3333     0.7777    3.00  0.00897 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.694 on 15 degrees of freedom
Multiple R-squared: 0.606,      Adjusted R-squared: 0.5534
 F-statistic: 11.53 on 2 and 15 DF,  p-value: 0.0009257
> model_2=lm(Hardness~x1+x2+x22)
> summary(model_2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   71.167      1.018  69.895  < 2e-16 ***
x1             2.917      0.720   4.051  0.00119 **
x2             2.333      0.720   3.241  0.00592 **
x22           -2.333      1.247  -1.871  0.08237 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.494 on 14 degrees of freedom
Multiple R-squared: 0.6848,     Adjusted R-squared: 0.6173
F-statistic: 10.14 on 3 and 14 DF,  p-value: 0.0008263
```

d) Which model will you suggest for the data? Explain your answer. Explain also why the estimates for the coefficients in front of x_1 and x_2 are the same in these two models.

Problem 3

Consider the linear model written in standard form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{X}$ an $n \times (p+1)$ matrix with rank $(p+1)$ and $\boldsymbol{\varepsilon}$ a vector of uncorrelated errors with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}$. Let $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t \boldsymbol{Y}$ be the least square estimator for $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

a) Find the mean vector and the covariance matrix of $\hat{\boldsymbol{\mu}}$.

The hat matrix $\boldsymbol{H}$ is defined as $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t$.

b) What are the properties of the $\boldsymbol{H}$ matrix that makes it possible to conclude that rank($\boldsymbol{H}$) = trace($\boldsymbol{H}$)? Show that $\sum_{i=1}^{n} Var\left(\hat{\mu}_i\right) = (p+1)\sigma^2$.

Now suppose you are fitting a model assuming expected response $E(Y) = \beta_0 + \beta_1 x_1$ while the true response is given by $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Let $\boldsymbol{H}$ be the hat matrix obtained assuming $E(Y) = \beta_0 + \beta_1 x_1$. Define $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}$.

c) Show that $\boldsymbol{e} = \beta_2 (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{x}_2 + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$.

Show that $E\left(\boldsymbol{e}^t \boldsymbol{e}\right) = (n-2)\sigma^2 + \beta_2^2 \boldsymbol{x}_2^t (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{x}_2$