Compulsory exercise 4

TMA4267 Linear statistical models 2020 spring

Magnus Wølneberg
Silje Anfindsen

April 16, 2020

Problem 1

a)

Does there exist a random vector having the following matrix as its covariance matrix?

Σ_1 : Yes. As we see that $Cov(X_i, X_j) = 0 \forall i \neq j$, we know that the components of the random vector $X = (X_1, X_2, X_3)$ are all independent.

Σ_2: No, not positive Semi-definitt as one of the eigenvalues are negative.

Σ_3: No, not symmetric, $\Sigma_3 \neq \Sigma_3^T$.

Σ_4: Yes, it is a vector of 3 dependent variables.

Σ_5: No, because X has to be equal to a constant. A random variable with no variance is not a random variable, hence it does not exist.

Σ_6: No, not symmetric $\Sigma_6 \neq \Sigma_6^T$.

Σ_7: No,not symmetric, $\Sigma_7 \neq \Sigma_7^T$.

Σ_8: No, not positive Semi-definitt as one of the eigenvalues are negative.

Σ_9: Yes, for example the vector $U = (X, X + Y, Y)$ as the first and third component are independent, and the second component are dependent on the other two.

b)

Do there exist two random vectors having the following matrix as their covariance matrix?

Σ_1: Yes, for example the two vectors $U = X$ and $V = \sigma X$.

Σ_2: Yes, for example $U = (X, Y, Z)$ and $V = (Z, Y, X)$.

Σ_3: Yes, for example $U = (1, X, 1)$ and $V = (X, 2X, 3X)$

Σ_4: Yes, here all the components of the two vectors are dependent. Example: $U = X$ and $V = 2X$

Σ_5: Yes, here all the components of the two vectors are independents. Example: $U = (X, Y, Z)$ and $V = (R, T, S)$

Σ_6: No, there does not exist two random vectors U and V satisfying: U_1, V_1 and U_1, V_2 dependent at the same time as U_2,V_2 and U_2, V_3 are dependent given that U_1 and V_3 are independent etc.

Problem 2

a)

Let $X = (X_1, X_2)^T$ be a bivariate random vector where both X_1 and X_2 are normally distributed. The vector $a^T X$ is not normal as long as the components X_1, X_2 are dependent. This is true for any bivariate vector a. An example of X is $(2Y, Y)$, where Y is normally distributed (and $2Y$ as this is a linear combination of a normal vector).

Under the additional condition that X_1 and X_2 are independent, there does **not** exist a vector $X = (X_1, X_2)$ and a such that $a^T X$ is not normally distributed.

b)

Uncorrelated variables are independent. The additional condition will therefore not change if we replace "uncorrelated" with "independent".

c)

Condition 1: X_1 and X_2 are normal.
Condition 2: (X_1, X_2) is normal.

The two conditions are not equivalent. Only condition 1 is satisifies. Since X_2 clearly is dependent on X_1, the bivariate distribution of X is not normal (see 2a)).

Problem 3

a)

The model fit can be measured by RSE (resiudal standard error) and R^2.

$RSE = 0.6168$.
This is the average amount that the response will deviate form the true regression line because of the error terms, ϵ_i. As the interval for y_i seems to be $(5, 81)$, the deviation of 0.62 is quite small in comparison.

$R^2 = 0.9993$
R^2 explains the proportion of variance explained by the model and in this case our model seems to explain most of the variance, which is good.

$$SSE = \sum_{i=1}^{15} (y_i - \hat{y_i})^2 = 4785,11$$

$$SST = \sum_{i=1}^{15}(y_i - \bar{y})^2 = 6685,73$$

$$SSR = \sum_{i=1}^{15}(\hat{y_i} - \bar{y})^2 = SST - SSE = 1900,62$$

b)

We know that

$$Var(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1} = SD(\hat{\beta})^2 = nSE^2, \tag{1}$$

where $SD(\hat{\beta})$ is the standard deviation for $\hat{\beta}$, and $SE = STD/\sqrt{n}$ is the standard error, and n the number of observations. We observe that $SE = (0.95841, 0.58319, 0.08128)^T$ in the R output.

We see that

$$\hat{\sigma} = \frac{1}{n-p}(Y - \hat{Y})^T(Y - \hat{Y}) = RSE^2, \tag{2}$$

and $RSE = 0.6168$ in the R output.

We can now derive the following result:

$$(X^TX)^{-1} = \frac{nSE^2}{\hat{\sigma^2}} = \frac{nSE^2}{RSE^2} = (36.22, 13.41, 0.26)^T \tag{3}$$

Problem 4

a)

The problem has the design matrix [100 x 3]:

$$X = \begin{pmatrix} 2 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 2 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ \vdots & \vdots & \vdots \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 2 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 2 \end{pmatrix} \tag{4}$$

where the first and last row with dots represent 25 rows, and the second row with dots represents 50 rows.

We have the vector

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T \tag{5}$$

Recall

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}, \tag{6}$$

and find an expression for this variance by doing matrix multiplication of X.

$$X^T X = 100I. \tag{7}$$

Now let

$$\hat{\beta}_1 = \underbrace{(1,0,0)}_{\text{A}}(\beta_1, \beta_2, \beta_3)^T, \hat{\beta}_2 = \underbrace{(0,1,0)}_{\text{B}}(\beta_1, \beta_2, \beta_3)^T, \hat{\beta}_3 = \underbrace{(0,0,1)}_{\text{C}}(\beta_1, \beta_2, \beta_3)^T \tag{8}$$

And, we finally see that

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(A\hat{\beta}, B\hat{\beta}) = A\sigma^2 (X^T X)^{-1} B^T = \frac{\sigma^2}{100}(1,0,0)(0,1,0)^T = 0.$$

We notice that the equation above gives the same result for any multiplication between each pair of the matrices A, B and C as these are linearly independent vectors, orthogonal to each other and therefore always zero when multiplied. Since the equation above is zero for all vectors of $\boldsymbol{\beta}$, independent of $\sigma(X^T X)^{-1}$, the result hold in general case.

Therefore we can say that $Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(\hat{\beta}_1, \hat{\beta}_3) = Cov(\hat{\beta}_3, \hat{\beta}_2) = 0$, and from the defintion, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are all independent.

Now we will look at the variance of the coefficents, given in (6):

Let $(X^T X)^{-1} = [c_{ij}]$ for $i, j = 1, 2, 3$ such that $\beta_j \sim N(\beta_j, \sigma^2 c_{jj})$ for $j = 1, 2, 3$. As we know that σ is equal for each β_j, and that $c_{11} = c_{22} = c_{33}$ from the result in (7), we get the following result:

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = Var(\hat{\beta}_3) = \frac{\sigma^2}{100}.$$

The result above only holds when all the diagonal elements in $(X^T X)^{-1}$ are equal. If for example the $x_{i2} = 2$ for $26 \leq i \leq 75$, the diagonal elements would differ.

b)

We are given this hypothesis:

$$H_0 : \beta_1 + \beta_3 = 2\beta_2$$

$$H_1 : \beta_1 + \beta_3 \neq 2\beta_2$$

We rewrite this into matrix form:

$$A\beta = d$$

where A is the matrix,

$$A = \begin{pmatrix} 1 & -2 & 1 \end{pmatrix} \tag{9}$$

and $d = 0$. We know that a general linar hypothesis test the test statistic is:

$$F = 1/r(A\hat{\beta} - d)^T(\hat{\sigma}^2 A(X^T X)^{-1}A^T)^{-1}(A\hat{\beta}). \tag{10}$$

We know that the unbiased estimator for $\hat{\sigma}^2$ is 3.27, $d = 0$, $A\hat{\beta} = 0.6516$ and $A(X^T X)^{-1}A^T)^{-1} = \frac{6}{100}$. Thus $F = 2.164$. We then look up in the statistical tables of critical values for $\alpha = 0.05$ fir the F-dist we find 3.94 for $r = 1$ and 97 degrees of freedom. Since

$$F_{OBS} < f_{1,97}$$

we will not reject the null hypothesis.

c)

We are given tree hypothesis tests;

$$H_0 : \beta_1 = 1 \text{ vs } H_1 : \beta_1 \neq 1$$

$$H_0 : \beta_1 + \beta_2 = 3 \text{ vs } H_1 : \beta_1 + \beta_2 \neq 3$$

$$H_0 : \beta_1 + \beta_2 + \beta_3 = 5 \text{ vs } H_1 : \beta_1 + \beta_2 + \beta_3 \neq 5$$

Since the hypothesis tests are dependent we use the Bonferroni method. The Bonferroni method gives us an upper bound to keep $FWER < 0.05$ and find α_{loc}:

$$\alpha_{loc} = \frac{\alpha}{m}$$

Then $\alpha_{loc} = 0.05, 0.025, 0.0167$ and we will only reject the last hypothesis test since $p = 0.0012 < \alpha_{loc} = 0.0167$.

Problem 5

a)

The defining relations of the fractional 2^{5-3} experiment with generators $D = AB$ and $E = AC$, are:

$I = ABD$ and $I = ACE$.

The resolution of the design is always the length of the shortes word in the defining relation, in this case it is 3.

Now, estimate the main factor A.

$$\begin{aligned}\hat{A} &= \text{expected mean response on high level} - \text{expected mean response on low level} \\ &= \frac{70,8+73,2+79,3+91,2}{4} - \frac{69,3+71,3+77,5+88.9}{4} = 1.875\end{aligned} \tag{11}$$

The estimated main effects of a factor measure the change in expected response when we move from low to high level (the factor changes two units).

We will now find the estimated effect of the interactions: BD, CE and $ABCDE$. In order to this, we find the sign for the interaction column by multiplying the signs in the columns for the main factors in the interaction. We then notice that the estimated effects for all the interactions above is 1.875

This means that the main factor A, and the interactions BD,CE and $ABCDE$ have the same importance in our design.