NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination: Mette Langaas

Phone: 988 47 649

Examination date: 4 June 2016

Examination time (from–to): 09:00–13:00

Permitted examination support material: C: Yellow, stamped A5 sheet with your own handwritten notes, Tabeller og formler i statistikk (Tapir forlag/Fagbokforlaget), K. Rottmann: Matematisk formelsamling. Specified calculator.

Language: English

Number of pages: 10

Number of pages enclosed: 0

Checked by:

Date Signature

Problem 1 Independent random variables

Assume that $\boldsymbol{X}$ is a bivariate normal random variable and that $\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ and $\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Let $\boldsymbol{Y} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \boldsymbol{X}$.

a) Find the distribution of $\boldsymbol{Y}$.

Specify a, b such that $\boldsymbol{Y}$ and $\begin{pmatrix} 2 & a \\ b & 1 \end{pmatrix} \boldsymbol{X}$ are independent random variables. Justify your answer.

Problem 2 Plant stress

At the Department of Biology at NTNU researchers use the model plant *Arabidopsis thaliana* to study the response of a plant to different sources of stress. In an experiment *Arabidopsis thaliana* seedlings were subject to a stress situation. The following factors were fitted:

- D (damage): $D = 1$ means that the plant was damaged mechanically by cutting into the leaves of the plant by a pair of scissors. $D = -1$ means damage was not inflicted (no cutting).
- F (flagellin): $F = 1$ means that the pathogen-derived peptide flagellin was sprayed on the leaves of the plant. $F = -1$ means water (not flagellin) was sprayed.
- T (time): Plants were harvested at two different time points after the stress situation. $T = 1$ means that the plant was harvested 60 minutes after the stress situation and $T = -1$ means that the plant was harvested 30 minutes after the stress situation.

Thus, we have three factors, D, F and T, each at two levels. In the study experiments for all possible combinations of the three factors were performed four times yielding 32 experiments in total.

The response measured in the experiment, was the observed gene activity level (a continuous measurement) of each of around 40 000 genes. We will only focus on the gene activity level of one of these genes, the AT1G32920 gene, and we denote the gene activity level of this gene by Y. It is known that this gene is active in response to wounding.

For experiment number i (where $i = 1, \ldots, 32$): Y_i is the observed response, D_i is chosen value of D, F_i is chosen value of F, and T_i is chosen value of T. A multiple regression model with all main effects, and two- and three-way interactions, was considered,

$$\begin{aligned} Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i \\ + \beta_{D:F} D_i F_i + \beta_{D:T} D_i T_i + \beta_{F:T} F_i T_i + \beta_{D:F:T} D_i F_i T_i + \varepsilon_i, \end{aligned}$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2. We refer to this as the *full model*. Note that the interactions are simply products of the factors. The vector of regression parameters is $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_D & \beta_F & \beta_T & \beta_{D:F} & \beta_{D:T} & \beta_{F:T} & \beta_{D:F:T} \end{pmatrix}^{\mathrm{T}}$, and the ith row of the design matrix $\boldsymbol{X}$ is $\begin{pmatrix} 1 & D_i & F_i & T_i & D_i F_i & D_i T_i & F_i T_i & D_i F_i T_i \end{pmatrix}$.

In Figure 1 you find R-commands and print-out from fitting the full model.

a) In the print-out from `summary(fit)` in Figure 1 *four* numerical values are replaced by question marks. Calculate numerical values for each of these, and explain what each of the values means.

A so called cube plot is given in the upper left panel of Figure 2. In the cube plot the fitted values from the multiple regression for the possible 8 combinations of the three factors are given. Plots of the main effects (upper right panel) and the interaction effects (lower panel) are found in Figure 2. In Figure 3 you find residual plots. See Figure 4 for the accompanying R-code and print-out.

b) How would you, based on Figures 2–4, evaluate the fit of the model?

How would you explain to a biologist what the estimated main effect of damage means in practice? How would you explain the estimated interaction effect between damage and flagellin?

Let $\gamma = 2^{\beta_F - \beta_D}$ be a new parameter of interest.

Suggest an estimator, $\hat{\gamma}$, for γ. Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$. Use results in Figure 1 to calculate numerical value for $\hat{\gamma}$, and estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

Hint: You may use that $2^x = \exp(x \ln 2)$, where ln is the natural logarithm.

The researchers want to test the hypothesis

$$\begin{aligned} &H_0 \colon \beta_{D:T} = \beta_{F:T} = \beta_{D:F:T} = 0 \qquad \text{vs.} \\ &H_1 \colon \text{at least one of } \beta_{D:T},\ \beta_{F:T},\ \beta_{D:F:T} \text{ is different from 0.} \end{aligned}$$

c) Perform the hypothesis test at a significance level of your own choice. All the numerical values you need for the calculations are found in Figure 1.

```
# data is in "standard order" in data frame with name "ds"
> ds %showing only rows 1-6 and 27-32 for space considerations
          Y  D  F  T
1  15.45169 -1 -1 -1
2  15.15908 -1 -1 -1
3  14.93064 -1 -1 -1
4  15.06569 -1 -1 -1
5  14.51032 -1 -1  1
6  14.76922 -1 -1  1
...
27 18.23645  1  1 -1
28 17.70327  1  1 -1
29 16.66523  1  1  1
30 16.96046  1  1  1
31 16.73133  1  1  1
32 16.57248  1  1  1
> fit=lm(Y~D*F*T,data=ds)
> model.matrix(fit)%only showing rows 1-6 and 27-32
   (Intercept)  D  F  T D:F D:T F:T D:F:T
1            1 -1 -1 -1   1   1   1    -1
2            1 -1 -1 -1   1   1   1    -1
3            1 -1 -1 -1   1   1   1    -1
4            1 -1 -1 -1   1   1   1    -1
5            1 -1 -1  1   1  -1  -1     1
6            1 -1 -1  1   1  -1  -1     1
...
27           1  1  1 -1   1  -1  -1    -1
28           1  1  1 -1   1  -1  -1    -1
29           1  1  1  1   1   1   1     1
30           1  1  1  1   1   1   1     1
31           1  1  1  1   1   1   1     1
32           1  1  1  1   1   1   1     1
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?       < 2e-16
D            0.93739    0.04140  22.644  < 2e-16
F            0.28546    0.04140   6.896 3.93e-07
T           -0.52354    0.04140 -12.647 4.18e-12
D:F         -0.08878    0.04140  -2.145  0.04231
D:T         -0.00242    ?        -0.058  0.95386
F:T         -0.12614    0.04140  -3.047  0.00555
D:F:T        0.09099    0.04140   2.198  ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

Figure 1: Printout from R-commands and statistical analyses for the plant stress data set. Four numbers are replaced by question marks.

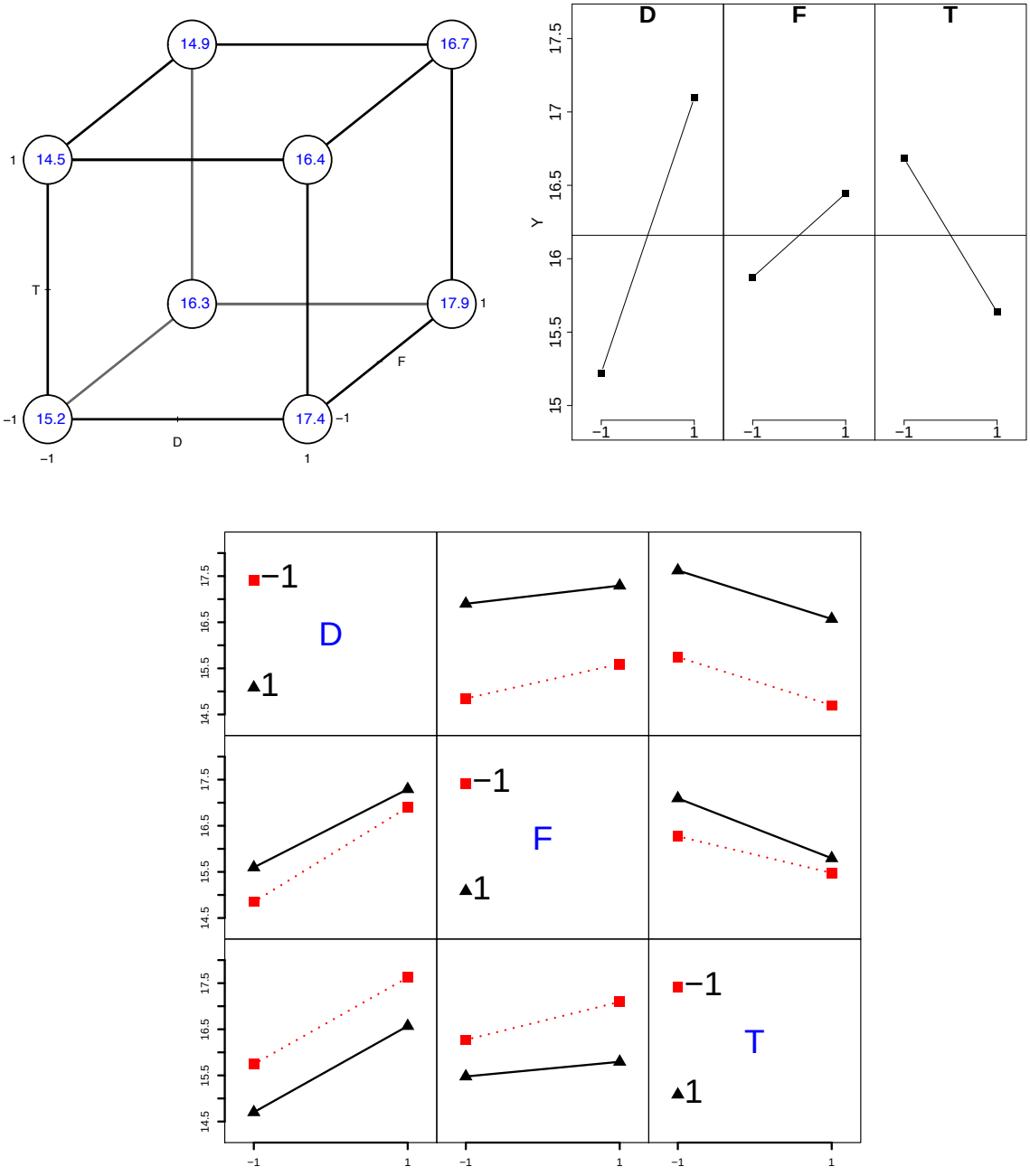

Figure 2: Cube plot (upper left panel), main effects plot (upper right panel) and interaction effects plot (lower panel) for the full model fitted to the plant stress data set.

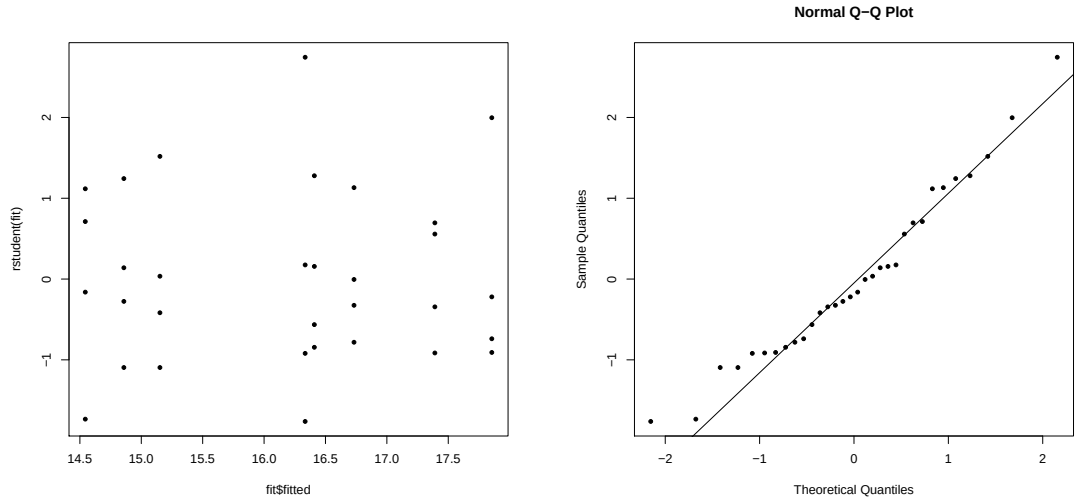

Figure 3: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for the full model fitted to the plant stress data set.

```
> library(FrF2)
> MEPlot(fit)
> IAPlot(fit)
> cubePlot(fit,"D","F","T",round=1,size=0.33,main="")
> plot(fit$fitted,rstudent(fit),pch=20)
> qqnorm(rstudent(fit),pch=20)
> qqline(rstudent(fit))
> ad.test(rstudent(fit))
        Anderson-Darling normality test
data:  rstudent(fit)
A = 0.43191, p-value = 0.2869
```

Figure 4: Print-out from R-commands and statistical analyses for the full model fitted to plant stress data set.

The researchers want to use the data to fit a prediction model, and want to consider reduced versions of the full model. First best subset model selection is used. Secondly, the researchers fit a lasso regression to the data. Results are presented in Figures 5 and 6.

d) Explain briefly what is done in the best subset model selection, and choose a good model based on the R^2_{adj}-criterion. Write down the fitted regression model for the model you choose.

Explain briefly what is done in the lasso regression, and write down the fitted regression model.

Compare the results from the best subset model selection and the lasso regression.

The researchers choose to use the following *reduced model* for prediction:

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \varepsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2. Output from fitting the reduced model is given in Figure 7.

e) Compare the estimated regression parameters and the estimated standard deviations of the estimated regression parameters for the full model (Figure 1) and the reduced model (Figure 7), and explain what you observe.

Based on the reduced model (Figure 7), provide a prediction and a 95% prediction interval for the gene activity level for the factor combination $D = 1,\ F = 1,\ T = -1$.

Hint: In a multiple linear regression with $n \times p$ design matrix $\boldsymbol{X}$, estimated regression coefficients $\hat{\boldsymbol{\beta}}$ and unbiased estimated error variance s^2, a $(1-\alpha)100\%$ prediction interval at $\boldsymbol{x}_0$ is given as

$$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2},n-p}\, s \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0},$$

where $t_{\alpha/2,n-p}$ denotes the value in the t-distribution with $n-p$ degrees of freedom that has area $\frac{\alpha}{2}$ to the right. See Figure 7 for some possible values for $t_{\alpha/2,n-p}$.

```
> x <- model.matrix(fit)[,-1]; dim(x)
[1] 32  7
> y <- ds$Y
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests=summary(bests)
> sumbests
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         D   F   T   D:F D:T F:T D:F:T
1  ( 1 ) "*" " " " " " " " " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " "*" " " "
5  ( 1 ) "*" "*" "*" " " " " "*" "*"
6  ( 1 ) "*" "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
> plot(bests,scale="adjr2",col=gray(seq(0.6,0.9,length=20)))
> round(sumbests$adjr2,3)
[1] 0.661 0.874 0.938 0.950 0.955 0.961 0.959
# LASSO
> library(glmnet)
> fit.lasso=glmnet(x,y,lambda=c(seq(1,0.01,length=60),0.001))
> plot(fit.lasso,xvar="lambda",label=TRUE)
> cv.lasso=cv.glmnet(x,y)
> log(cv.lasso$lambda[which.min(cv.lasso$cvm)])
[1] -4.716347
> coef(cv.lasso,s="lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) 16.15941869
D            0.92843876
F            0.27651524
T           -0.51459006
D:F         -0.07983558
D:T          .
F:T         -0.11719275
D:F:T        0.08204094
```

Figure 5: Print-out from R performing best subset selection and lasso regression on the plant stress data.

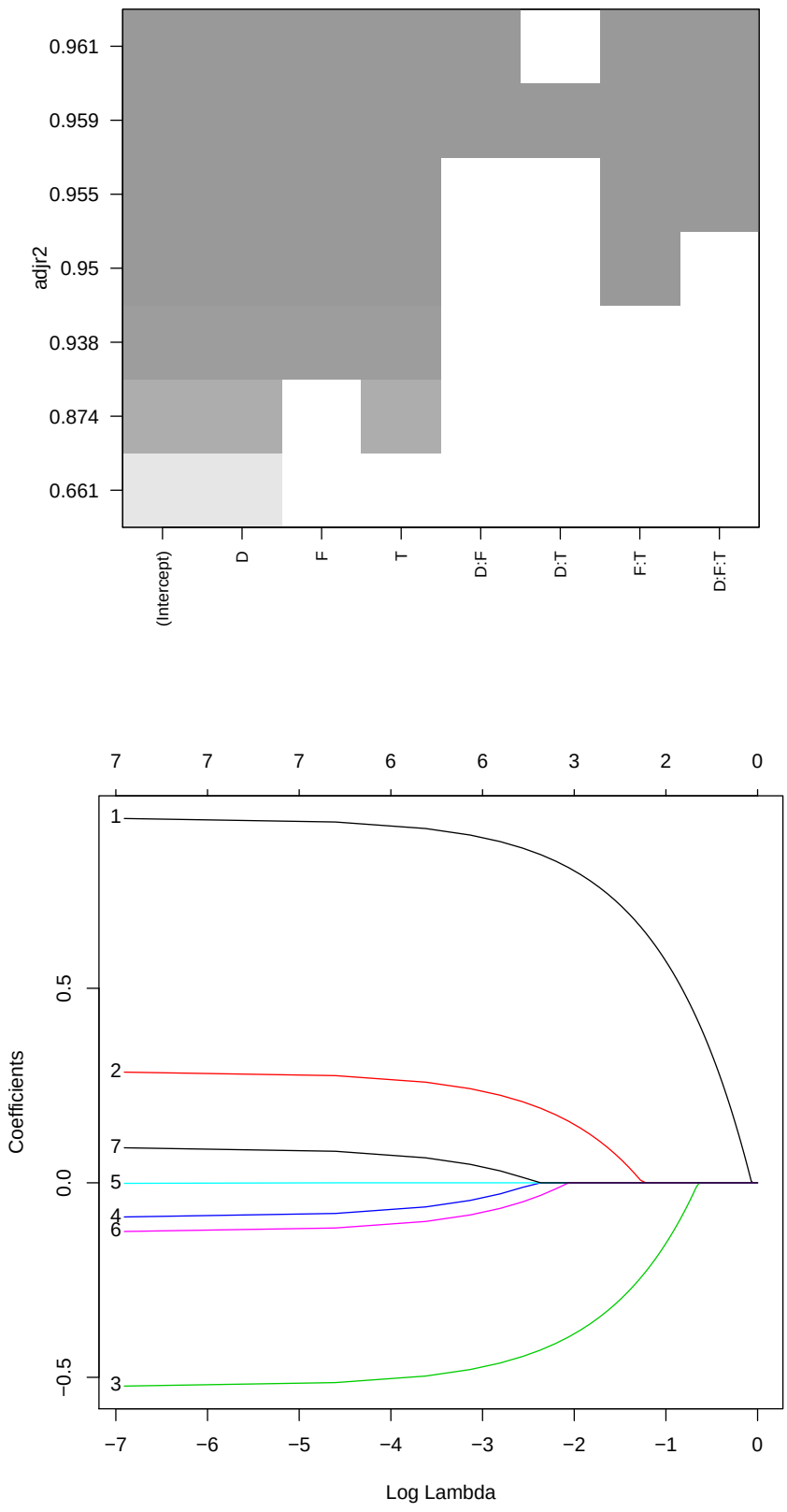

Figure 6: Figure (upper panel) from model selection and lasso regression (lower panel), with R-code in Figure 5. Coding for lines in lasso figure is 1=$\hat{\beta}_D$, 2=$\hat{\beta}_F$, 3=$\hat{\beta}_T$, 4=$\hat{\beta}_{D:F}$, 5=$\hat{\beta}_{D:T}$, 6= $\hat{\beta}_{F:T}$ and 7=$\hat{\beta}_{D:F:T}$.

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:   0.95,        Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

Figure 7: Print-out from R performing linear regression on the reduced model for the plant stress data set.

Problem 3 Properties of estimator for σ^2

Let $\boldsymbol{Y}$ be an $n \times 1$ random vector with mean $\mu\mathbf{1}$ and covariance matrix $\sigma^2\boldsymbol{I}$, where $\mathbf{1}$ is an $n \times 1$ vector with all elements equal to 1 and $\boldsymbol{I}$ is an $n \times n$ identity matrix. Further, denote by Y_i element i of $\boldsymbol{Y}$, and let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$.

An estimator for σ^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}\Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Big)\boldsymbol{Y}.$$

We give the following useful result. Let $\boldsymbol{X}$ be an $n \times 1$ random vector with mean $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma}$, and let $\boldsymbol{C}$ be an $n \times n$ symmetric constant matrix. Then,

$$\mathrm{E}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\eta}. \tag{1}$$

a) First, write down the value of $\mathbf{1}^{\mathrm{T}}\mathbf{1}$, and the matrices $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ and $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ for $n = 4$.
What are key characteristics of the matrix $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ (symmetric or not, idempotent or not, rank)?
Use Equation (1) to find $\mathrm{E}(S^2)$.

Let us now assume that $\boldsymbol{Y}$ is multivariate normally distributed with the mean and covariance given above.

b) Show that $\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$ follows a χ^2-distribution, and also derive the number of degrees of freedom.
Use this result to find the variance of S^2.
Is the random variable $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$ and the random vector $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$ independent? Justify your answer.
Finally, find the distribution of

$$\frac{n(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}}.$$

Justify your answer.