NTNU
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination: Øyvind Bakke

Phone: 73 59 81 26, 990 41 673

Examination date: 3 June 2019

Examination time (from–to): 9:00–13:00

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

Other information:
In the grading, each of the eight points counts equally. All answers must be justified, and relevant calculations provided.

Language: English

Number of pages: 4

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☐ **2-sidig** ☒
sort/hvit ☒ **farger** ☐
skal ha flervalgskjema ☐

Date Signature

Problem 1

Assume that $\boldsymbol{X} = (X_1 \ X_2)^{\mathrm{T}}$ has a bivariate normal distribution with covariance matrix

$$\Sigma_X = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix},$$

with a a real number.

a) What do we require of a covariance matrix of a random vector? For which a is Σ_X a covariance matrix?

Assume that $\boldsymbol{Y} = (Y_1 \ Y_2 \ Y_3)^{\mathrm{T}}$ has a trivariate normal distribution with covariance matrix

$$\Sigma_Y = \begin{pmatrix} 1 & a & 0 \\ a & 1 & b \\ 0 & b & 1 \end{pmatrix},$$

with a and b real numbers.

b) First we return to $\boldsymbol{X}$: What is the covariance of $X_1 + X_2$ and $X_1 - X_2$? For which a are the two independent?

For which a and b are $Y_1 + Y_2 + Y_3$ and $Y_1 - Y_2 - Y_3$ independent (and Σ_Y is a covariance matrix)?

Problem 2

Suppose you want to run a 2^{5-2} fractional factorial experiment and have chosen $D = AB$ and $E = AC$ as generators for the design.

What is the resolution of a fractional factorial experiment? Why do we want it as high as possible? What is the resolution of the above experiment? Are any main effects aliased with any 2-factor interaction? If yes, which?

Problem 3

The taste of 30 samples of cheddar cheese was studied. A multiple linear regression model was fitted, with acetic acid (Norwegian: *eddiksyre*) concentration (`acetic`), lactic acid (*melkesyre*) concentration (`lactic`) and the logarithm of hydrogen sulfide concentration (`logh2s`) as covariates. The response was a taste score (`taste`) made by judges. (Data from Dunn and Smyth, *Generalized linear models with examples in R.*) R input and output and some plots are shown in Figure 1.

a) Explain how an original model was reduced using best subset selection. Comment briefly on the model fit of the reduced model. Calculate the error sum of squares (SSE, also called residual sum of squares) of the reduced model.

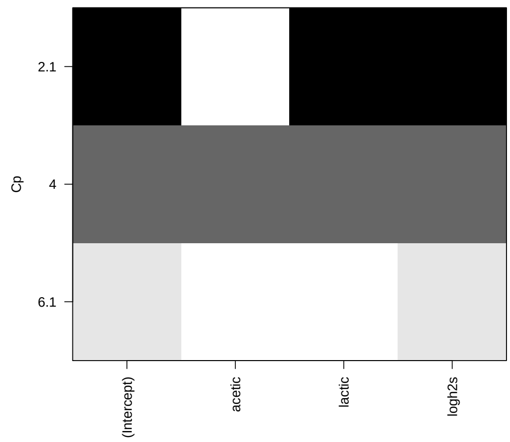
We have seen that the covariance matrix of the coefficient estimators in a linear regression model is $\sigma^2(X^{\mathrm{T}}X)^{-1}$, with σ^2 the variance of the errors and X the design (model) matrix. In a model with intercept, it can be shown that this gives the variance

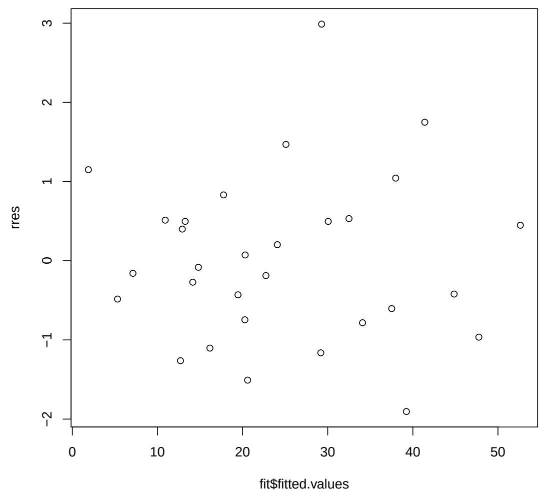
$$\operatorname{Var}\hat{\beta}_j = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2},$$

of a coefficient estimator $\hat{\beta}_j$ (for a coefficient that is not the intercept). Here, the x_{ij} are the n values of covariate j and $\bar{x}_j$ their mean, and R_j^2 the coefficient of determination (multiple R^2) for the regression with x_j as response and all the other covariates of the original model as covariates.

b) Discuss conditions that will lead to high or low variance of $\hat{\beta}_j$.

```
> cheesedata<-data.frame(taste,acetic,lactic,logh2s)
> summary(lm(taste~.,data=cheesedata))

Call:
lm(formula = taste ~ ., data = cheesedata)

Residuals:
     Min       1Q   Median       3Q      Max
-17.5250  -6.6580  -0.8226   5.0833  24.9859

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.142493   9.277924  -2.925  0.00705 **
acetic        0.004184   0.014916   0.281  0.78129
lactic       19.201965   8.457616   2.270  0.03171 *
logh2s        3.836799   1.219895   3.145  0.00413 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.12 on 26 degrees of freedom
Multiple R-squared:  0.6528,Adjusted R-squared:  0.6127
F-statistic: 16.29 on 3 and 26 DF,  p-value: 3.675e-06

> library(leaps)
> best<-regsubsets(taste~.,data=cheesedata)
> summary(best)$which
  (Intercept) acetic lactic logh2s
1        TRUE  FALSE  FALSE   TRUE
2        TRUE  FALSE   TRUE   TRUE
3        TRUE   TRUE   TRUE   TRUE
> summary(best)$cp
[1] 6.108163 2.078697 4.000000
> plot(best,scale="Cp")
> fit<-lm(taste~lactic+logh2s)
> summary(fit)

Call:
lm(formula = taste ~ lactic + logh2s)

Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.529  -1.163   4.844  25.617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.591      8.982  -3.072  0.00481 **
lactic        19.886      7.959   2.498  0.01886 *
logh2s         3.946      1.136   3.475  0.00174 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared:  0.6517,Adjusted R-squared:  0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07

> rres<-rstudent(fit)
> plot(fit$fitted.values,rres)
> qqnorm(rres)
> qqline(rres)
```





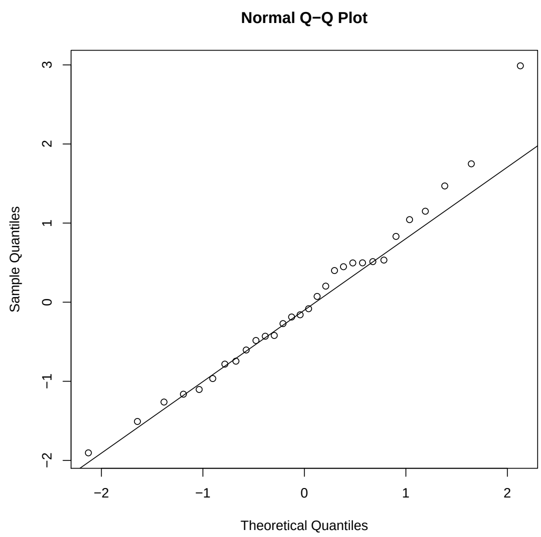


Figure 1: Model from Problem 3a: R input and output (left), a graphical table of best subsets using Mallows' C_P as the statistic for ordering models (upper right), residual plot of reduced model (middle right), normal Q–Q plot of reduced model (lower right). Note that the information of the graphical table is also included in the R output.

Problem 4

A response variable Y_{kj} was measured, using 10 repetitions for each of three levels $j = 1, 2, 3$ of a factor. A regression model of the form $Y_{kj} = \mu_j + \epsilon_{kj}$ was assumed, where $k = 1, 2, \ldots, 10$, and the ϵ_{kj} were independent $N(0, \sigma^2)$. Then it is given that the design matrix of the model has dimensions 30×3 and that $X^{\mathrm{T}}X = 10I$, with I a 3×3 identity matrix.

We want to perform pairwise comparisons, i.e., perform three hypothesis tests, in which the null hypotheses are

$$\mu_1 = \mu_2, \qquad \mu_1 = \mu_3, \qquad \mu_2 = \mu_3,$$

respectively, against two-sided alternatives.

The least-squares estimates of μ_2 and μ_3 were 0.2488 and 1.1663, respectively, and the error sum of squares was SSE = 24.00.

a) Perform the test in which the null hypothesis is $\mu_2 = \mu_3$. Use significance level 0.05. You should calculate a test statistic and use its distribution under the null hypothesis to arrive at your conclusion.

A corresponding test was performed for all pairs of coefficients. The p-values are given in the following table.

Null hypothesis:	$\mu_1 = \mu_2$	$\mu_1 = \mu_3$	$\mu_2 = \mu_3$
p-value:	0.784	0.021	0.038

b) What is *Bonferroni's method* for family-wise error rate (FWER) control? Which null hypotheses are rejected if Bonferroni's method is used to keep the FWER below 0.05 when performing the above tests?

Consider a different method for performing the three tests: The null hypothesis is rejected if it is rejected at the 0.05 significance level *and* in addition the null hypothesis $\mu_1 = \mu_2 = \mu_3$ against the alternative that at least one differs from the others, is also rejected at the 0.05 significance level.

c) Show that this method will also keep the FWER below 0.05. (Hint: Consider the different combinations of the three null hypotheses being true and false.)

The p-value of the test of $\mu_1 = \mu_2 = \mu_3$ is 0.041.

Which of the three null hypotheses would be rejected by this method?