

i Front page

Examination paper for TMA4257 Linear Statistical Models

Date: 06.05.2026

Time: 15:00 - 19:00

Course contact: Thea Bjørnland

Present at the exam location: NO

Permitted examination support material:

C:

- Tabeller og formel i statistikk, Tapir forlag
- Basic calculator

OTHER INFORMATION

Read the questions carefully and make your own assumptions. In your answers, explain clearly what assumptions you have made and how you have understood or limited the assignment.

If there are direct errors or omissions in the assignment set and you cannot make your own assumptions, please refer to the information about complaints regarding formal errors on the NTNU website "Explanation of grades and appeals".

SPECIFIC INFORMATION FOR YOUR COURSE

For the entire exam you are meant to answer on handwritten sheets. At the bottom of the question page, you will find a seven-digit code. Fill in this code in the top left corner of the sheets you wish to submit.

Please note that only blue or black ballpoint pens are allowed.

We recommend that you do this during the exam. If you require access to the code after the examination time ends, click "Show submission".

You are responsible for filling in the correct code on the handwritten sheets. Therefore, read the cover sheet carefully. The Examination Office cannot guarantee that incorrectly completed sheets will be added to your assignment.

Weighting: The exam consists of 10 subproblems, which contribute equally to the final grade.

Withdrawing from the exam:

If you wish to submit a blank test/withdraw from the exam for another reason, go to the "hamburger menu" in the top right-hand corner and click "Submit blank". This cannot be undone, even if the test is still open.

Access to your answers:

After the exam, you can find your answers under previous tests in Inspira. Be aware that it may take a working day until any paper drawings material is available in "previous tests".

1 Problems

All questions (10 sub-problems) are found in the PDF and should be answered with pen and paper.

Maximum marks: 10

Question 1

Attached



Problem 1

Let $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$ be a multivariate normal random vector with mean and covariance

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

(a) What is the distribution (including parameter values) of Y_2 ?

Show that the expected value $E(Y_2 \mid Y_1 = y_1)$ can be written on the form $\beta_0 + \beta_1 y_1$, and find the values of β_0 and β_1 .

(b) Let $Z_1 = Y_1 + Y_2 + Y_3$ and $Z_2 = 3Y_1 + Y_2 - Y_3$ and $\mathbf{Z} = (Z_1, Z_2)^\top$. What is the distribution of $\mathbf{Z}$? Include calculations of parameter values.

Problem 2

An experiment was conducted to investigate the effect of levels two factors (A and B) on some response Y . Equally many experiments were conducted at each combination of factor levels. The resulting ANOVA table is given below.

(a) How many levels of the factors A and B were considered? How many experiments were conducted at each combination of factor levels?

Fill in the three missing parts (marked with ?) of the table.

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	162.86	162.86	8.3205	0.006581 **
B	1	518.62	518.62	26.4959	9.568e-06 ***
A:B	?	23.46	23.46	?	0.280926
Residuals	36	704.64	?		

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

Problem 3

Consider $\mathbf{Y} \sim \text{MVN}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{X}$ is an $n \times p$ matrix of rank $p < n$.

- (a) Show that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is idempotent with rank p .
- (b) Find $E(\mathbf{Y}^\top\mathbf{H}\mathbf{Y})$ in terms of $\mathbf{X}$, $\boldsymbol{\beta}$, σ^2 and p . Find the two parameters in the distribution of $\frac{1}{\sigma^2}\mathbf{Y}^\top\mathbf{H}\mathbf{Y}$.
- (c) Find the distribution of

$$Q = \frac{\mathbf{Y}^\top\mathbf{H}\mathbf{Y}/p}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}/(n - p)}.$$

Problem 4

Consider the model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{V})$, $\mathbf{V}$ is symmetric, positive definite, and known, and $\boldsymbol{\beta}_2 \neq \mathbf{0}$.

- (a) Which, if any, requirements for $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{V}$ are necessary for the estimator

$$\tilde{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{Y}$$

to be unbiased for $\boldsymbol{\beta}_1$? Justify your answer.

Problem 5

A (fictitious) grocery store chain is planning to build a new store. Your job is to help them understand how weekly turnover is related to key store characteristics, using data from $n = 25$ existing stores. For each store, you have the following variables:

- **turnover**: Weekly turnover (in 1000 NOK).
- **population**: Number of people living within 1 km of the store.
- **distance**: Distance (in km) to the nearest competing grocery store.
- **premium_location**: Indicator; 1 if high-income residential area, otherwise 0.

You regress turnover against the covariates population, distance and premium_location. The model summary is presented below.

Call:

```
lm(formula = turnover ~ distance + population + premium_location)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-263.936	-70.211	-1.167	83.698	313.881

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	268.04461	100.05773	2.679	0.01405 *
distance	68.12934	20.45226	3.331	0.00317 **
population	0.22219	0.03628	6.125	4.46e-06 ***
premium_location	78.01698	56.72120	1.375	0.18349

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

Residual standard error: 138.7 on 21 degrees of freedom

Multiple R-squared: 0.7602, Adjusted R-squared: 0.726

F-statistic: 22.19 on 3 and 21 DF, p-value: 1.03e-06

(a) Calculate a 95% confidence interval for the coefficient of `premium_location`. Does being in a premium location appear to increase turnover?

(b) Using the the plots in Figure 1 (page 4), comment on the relevant model assumptions.

For multiple linear regression models in general, show that when the model assumptions hold, the residuals and fitted values (predictions) are uncorrelated.

(c) To improve your model, each store is now observed for 8 weeks, giving 25 stores $\times$ 8 weeks = 200 observations of weekly turnover. Which assumption of the ordinary linear regression model is likely to be violated if applied to this data set?

Now, we apply the linear mixed model

$$\text{turnover}_{ij} = \beta_0 + \beta_1 \text{distance}_{ij} + \beta_2 \text{population}_{ij} + \beta_3 \text{premium_location}_{ij} + \gamma_i + \varepsilon_{ij},$$

where $i = 1, \dots, 25$ and $j = 1, \dots, 8$, and $\gamma_i \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $\gamma_i, \varepsilon_{ij}$ are independent. Estimates of variance components are $\hat{\tau}^2 = 12126$ and $\hat{\sigma}^2 = 4759$. What does $\frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \approx 0.72$ represent? Justify your answer.

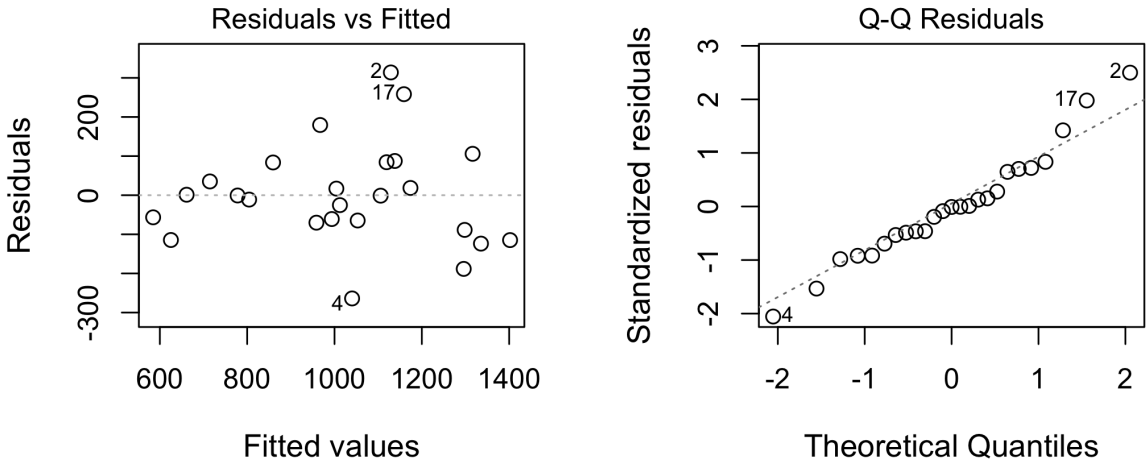


Figure 1: Residual plots