

i Front Page

Examination paper for ST2304 Statistical Modelling for Biologists/Biotechnologists

Date: 2025-05-13

Time: 09:00 - 13:00

Course contact: Bob O'Hara

Present at the exam location: NO

Permitted examination support material: **C** One hand-written A4 sheet of paper is allowed.. A specific basic calculator is allowed, and mathematical tables (Tabeller og formel i statistikk (Tapir forlag, Fagbokforlaget), Matematiske formelsamling (K. Rottmann)).

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

No paper drawings: This exam does not include hand drawings. If you receive hand drawing sheets, this is by mistake. **You will not be able to submit the sheets, and they will not be graded.**

Weighting: The maximum achievable score for each question is indicated on the question.

Withdrawing from the exam:

If you wish to submit a blank test/withdraw from the exam for another reason, go to the 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 hand-written material is available in "previous tests".

1 Multiple Choice: Linear model

Which of the following are standard assumptions of a classical linear regression model (e.g., Ordinary Least Squares - OLS)?

Select one or more alternatives:

- ☐ The relationship between the independent and dependent variable is linear.
- ☐ The residuals are zero-mean normally distributed.
- ☐ The sample size must be greater than 100.
- ☐ The independent variables are normally distributed
- ☐ The residuals have constant variance (homoscedasticity).
- ☐ The independent variables are measured without error.

Maximum marks: 3

2 Are you confident about confidence intervals?

Which of the following statements about confidence intervals are generally true?
(Select all that apply)

Select one or more alternatives:

- ☐ If the confidence interval for a mean difference includes zero, there may be no significant difference between the groups.
- ☐ A wider confidence interval indicates more precise estimates of the population parameter.
- ☐ Increasing the sample size generally results in a narrower confidence interval.
- ☐ A 95% confidence interval means there is a 95% probability that the population parameter lies within the interval.
- ☐ A 95% confidence interval will contain the sample mean 95% of the time.

Maximum marks: 2

A team of plant ecologists is investigating how **leaf nitrogen content** and **sunlight exposure** affect the **photosynthetic rate** of a plant species. They collect data on:

- Photo: Photosynthesis rate ($\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$)
- N: Leaf nitrogen content (mg/g)
- Light: Sunlight exposure (hours per day)

They use R to analyze the data and obtain the following output:

```
Call:
lm(formula = Photo ~ N + Light, data = plant_data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.14538 -0.47645 -0.01234  0.47367  2.38427

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9853     0.5674   3.50  0.0010 **
N              0.4876     0.0923   5.28 2.0e-06 ***
Light          0.2387     0.0698   3.42  0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.812 on 47 degrees of freedom
Multiple R-squared:  0.624, Adjusted R-squared:  0.608
F-statistic: 39.0 on 2 and 47 DF, p-value: 3.2e-10
```

3 What is the Equation?

Write out the full regression equation, based on the output

Fill in your answer here

Format

B

I

U

x_2

x^2

I_x

Words: 0

Maximum marks: 4

4 What Do the Coefficients Mean?

Interpret the coefficients of both N and $Light$ in context.

Fill in your answer here

Format

B

I

U

x_2

x^2

I_x

Words: 0

Maximum marks: 4

5 R squared?

What does the R-squared value tell you in this model?

Fill in your answer here

Format

B *I* U x_2 x^2 I_x

    

   

        

Words: 0

Maximum marks: 2

The figure displays three diagnostic plots for a linear regression model, arranged vertically.

- Residual Plot:** The y-axis is labeled "Residuals" and ranges from -8 to 2. The x-axis is labeled "Fitted Values" and ranges from -25 to 15. The plot shows a scatter of residuals around a horizontal line at zero. Most points are within the range of -4 to 2, but there are two notable outliers at approximately (-25, -8) and (15, -8).
- Normal Q-Q Plot:** The y-axis is labeled "Sample Quantiles" and ranges from -8 to 2. The x-axis is labeled "Theoretical Quantiles" and ranges from -2.5 to 2.5. The plot shows a line of best fit and a scatter of points. The points generally follow the line, but there are two significant outliers at approximately (-2.5, -8) and (2.5, -8).
- Cook's D:** The y-axis is labeled "Cook's D" and ranges from 0.0 to 2.0. The x-axis is labeled "Fitted Values" and ranges from -25 to 15. The plot shows a scatter of Cook's D values. Most points are near zero, but there are two significant outliers at approximately (-25, 2.5) and (15, 0.8).

Fill in your answer here

Format ▾ | **B** *I* U x₂ x² | T_x | 📄 📋 | ⬅ ➡ ↺ | ⅞ :≡ ±≡ ∓≡ |

☰ ☷ ≡ ≡ | Ω 🗃️ | ✎ | Σ | ✖

Words: 0

Maximum marks: 6

7 Improving the Model

Suggest two ways you could try to improve the model to mitigate the problems you have identified.

Fill in your answer here

Format

B

I

U

x_2

x^2

I_x





































Words: 0

Maximum marks: 4

In the regression model the two covariates (leaf nitrogen content and sunlight exposure) are not standardized. They are used in their original units (e.g., mg/g for nitrogen content and hours per day for sunlight exposure).

8 What is Standardisation?

What do we mean by standardisation and when would it be useful to standardise the covariates?

Fill in your answer here

Maximum marks: 4

9 Standardised Interpretations

How might the interpretation of the estimated covariate coefficients change if you standardised the covariates? (you would need to know more about the covariates, so feel free to illustrate with suggestions like "if the covariates looked like this then...")

Fill in your answer here

Maximum marks: 4

A wildlife biologist is studying the daily survival of a small mammal species in a protected forest. Over a 10-day period, she tracks 12 tagged individuals. Each day, she records whether each animal is alive (1) or dead (0). The results are summarized as follows:

- Total number of animal-days observed: 120
- Total number of survival events (animal was alive the next day): 108

Assume survival follows a Bernoulli process (each day's survival is independent and has the same probability p).

10 Estimated Survival

Compute the maximum likelihood estimator for the daily survival probability $\hat{p}$:

Maximum marks: 1

11 Confidence Intervals

Compute a 95% confidence interval for p using a normal approximation, to 2 decimal places.

The lower bound is .

The upper bound is .

Maximum marks: 4

A biologist is studying how habitat type, plant height, and their interaction influence the number of herbivorous insects found on plants. Data were collected on 150 plants, including:

- Count: Number of insects on each plant (count data).
- Habitat: "wet" or "dry"
- Height: Height of plant in cm (numeric, centered to have a mean of zero).

The biologist fits two models:

```
model1 <- glm(N ~ Habitat + Height, family="poisson",
data=insect_data)
```

```
model2 <- glm(N ~ Habitat * Height, family="poisson",
data=insect_data)
```

and obtains the following summary information for model2:

Call:

```
glm(formula = N ~ Habitat * Height, family = "poisson", data = insect_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.968345	0.073864	13.110	< 2e-16 ***
Habitatwet	0.651597	0.090852	7.172	7.39e-13 ***
Height	0.009988	0.001468	6.804	1.02e-11 ***
Habitatwet:Height	-0.002655	0.001737	-1.529	0.126

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 191.722 on 149 degrees of freedom

Residual deviance: 16.302 on 146 degrees of freedom

AIC: 489.49

Number of Fisher Scoring iterations: 4

12 What Link?

What link function do you think was used?

Select one alternative:

- ☐ probit
- ☐ log
- ☐ identity
- ☐ logit
- ☐ cloglog

Maximum marks: 1

13 How Many Insects?

From model2, what is the expected number of insects on a dry-habitat plant that is 10cm taller than the average (answer to one decimal place)?

Please fill in: .

From model2, what is the expected number of insects on a wet-habitat plant that is 10cm taller than the average (answer to one decimal place)?

Please fill in: .

Maximum marks: 6

2	146	16.302	1	2.3589	0.1246
---	-----	--------	---	--------	--------

14/15

- What does this indicate about the fit of the model?
- What are two ways the biologist could improve the model to account for this?

[illegible]

15/15