STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 1 of 2

Submission deadline

Thursday October 2 2025, 14:30 in Canvas (`canvas.uio.no`).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts. You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with Latex). The assignment must be submitted as **a single PDF file**. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons,you have to contact the person responsible for the course, Ingrid Hobæk Haff (e-mail: ingrihaf@math.uio.no) no later than the same day as the deadline. All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Specifically about this assignment

In order to get the assignment accepted you need to fulfil the following requirements:

- Made a real attempt on all (sub-)questions
- Give satisfactory answers in at least 60% of the (sub-)questions
- Include relevant code outputs in your report.

Complete guidelines about delivery of mandatory assignments:

`www.uio.no/english/studies/examinations/compulsory-activities/mn-math-mandatory.html`

GOOD LUCK!

Problem 1

In this problem, we will consider the relationship between a person's wingspan and height. The wingspan is the horizontal measurement from fingertip to fingertip with outstretched arms. The data below show the wingspan and height in cm for 16 Australian women, and are also recorded in the file `http://www.uio.no/studier/emner/matnat/math/STK3100/data/wingspan.txt`:

	Height (x)	Wingspan (y)
1	63.0	62.0
2	63.0	62.0
3	65.0	64.0
4	64.0	64.5
5	68.0	67.0
6	69.0	69.0
7	71.0	70.0
8	68.0	72.0
9	68.0	70.0
10	72.0	72.0
11	73.0	73.0
12	73.5	75.0
13	70.0	71.0
14	70.0	70.0
15	72.0	76.0
16	74.0	76.5

We will return to the relationship between height and wingspan in question f), but first we will consider the problem more generally. To this end we consider a simple linear regression model with the single covariate $\mathbf{x} = (x_1, \ldots, x_n)^T$ and the response $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma^2)$.

We will consider two models, M_0 and M_1. Model M_1 is the standard linear regression model

$$\mu_i = \beta_0 + \beta_1 x_i,$$

whereas M_0 is the same model, but without the intercept, i.e.

$$\mu_i = \beta_1^* x_i.$$

Now, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ be the $n \times 1$ vector of mean values. Further, the model matrices for models M_0 and M_1 are denoted $\mathbf{X}_0$ and $\mathbf{X}_1$ and the model

spaces are denoted $C(\mathbf{X}_0)$ and $C(\mathbf{X}_1)$. We use the notation $\mathbf{1}_k$ to denote a $k \times 1$ vector of 1s.

a) Give the model matrices for models M_0 and M_1. What are the ranks of the two model matrices? Explain what it means that the models are nested.

b) Give the projection matrices $\mathbf{P}_0$ and $\mathbf{P}_1$ onto the two model spaces.

c) Use the projection matrices to show that the vectors of fitted values for models M_0 and M_1 may be given, respectively, as

$$\hat{\boldsymbol{\mu}}_0 = \hat{\beta}_1^* \mathbf{x}$$

and

$$\hat{\boldsymbol{\mu}}_1 = \bar{Y}\mathbf{1}_n + \hat{\beta}_1(\mathbf{x} - \bar{x}\mathbf{1}_n),$$

with

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

d) Show that

$$\mathbf{Y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}/\sigma^2 \quad \text{and} \quad \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{Y}/\sigma^2$$

are independent and determine their distributions (Hint: Use Cochran's theorem). It is sufficient to determine the distributions under M_0.

e) Show that the F -statistic for testing the null hypothesis that model M_0 holds versus the alternative hypothesis that model M_1 holds may be given as

$$F = \frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2}{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}_1\|^2/(n-2)} = \frac{\sum_{i=1}^n \left(\bar{Y} - \hat{\beta}_1 \bar{x} + \left(\hat{\beta}_1 - \hat{\beta}_1^*\right) x_i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y} - \hat{\beta}_1(x_i - \bar{x})\right)^2 /(n-2)},$$

and determine the distribution of the F-statistic under model M_0.

We now return to the example on wingspan and height, considered in the beginning of the problem.

f) Read the data in the table given in the beginning of the problem into R. Use the `lm` command to fit models M_0 and M_1, and use the `anova` command to perform the F test. Discuss your results.

Problem 2

In this problem, we will consider how the survival of a passenger on the Titanic depended on ticket class and the passenger's age. The file

`https://www.uio.no/studier/emner/matnat/math/STK3100/data/titanic.txt`

consists of three columns with the following information about a subset of 70 passengers:

- `survived`: survived the shipwreck (0 = no, 1 = yes)
- `age`: age of the passenger in years
- `pclass`: ticket class, grouped as either 3rd class or 1st/2nd class (0 = 1st/2nd class, 1 = 3rd class)

We will use logistic regression to analyse the data using R. You may read the data into R by the commands:

```
data="http://www.uio.no/studier/emner/matnat/math/STK3100/data/titanic.txt"
titanic=read.table(data,header=T)
```

a) Fit a logistic regression model with `survived` as response and `pclass` as the only covariate. Is there a significant effect of `pclass`?

b) Denote by β_1 the regression coefficient for `pclass` in the logistic regression model. Give an interpretation of e^{β_1}, and estimate it.

c) Fit a logistic regression model where you also include `age` as a covariate. Denote by β_2 the regression coefficient for `age` in this model. Give an interpretation of e^{β_2}, and estimate it.

d) Use the Wald test, the likelihood ratio test and the score test to test the hypothesis that $\beta_2 = 0$ in the model in question c). How well do the tests agree? What are the conclusions of the tests?

e) Find 95% confidence intervals for e^{β_1} from b) and for e^{β_2} from c). Give interpretations of these intervals.