14th September, 2023

STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 1 of 2

Submission deadline

Thursday 28th September 2023, 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 Student Administration at the Department of Mathematics (e-mail: studieinfo@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.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

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

- Made a real attempt on all (sub-)questions (except 3 e)).
- Give satisfactory answers in at least 60% of the (sub-)questions (not counting 3 e))
- Include relevant R outputs in your report

Problem 1. The situation we will consider in this problem may be illustrated by the following study. A biologist wished to study the effects of ethanol on sleep time. A sample of 20 rats, matched for age and other characteristics, was selected, and each rat was given an oral injection having a particular concentration of ethanol per kg of body weight. The rapid eye movement (REM) sleep time for each rat was then recorded for a 24 hour period. The aim of the study was to investigate if and how the amount of ethanol affects the time of REM sleep.

The results of the study are given in the table below.

Amount ethanol (g/kg)	REM time				
0	88.6	73.2	91.4	68.0	75.2
1	63.0	53.9	69.2	50.1	71.5
2	44.9	59.5	40.2	56.3	38.7
4	31.0	39.6	45.3	25.2	22.7

We will return to the study of REM sleep and ethanol in question f. First we will consider the problem more generally. To this end we consider a one-way layout with c groups and observations $Y_{i1}, \ldots, Y_{in_i}$ for the n_i individuals in group i. We assume that all observations are independent and that $Y_{ij} \sim N(\mu_i, \sigma^2)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, c$. We also assume that there is a numeric quantity x_i that gives the "exposure" to some substance for the individuals in group i (like the amount of ethanol in the example).

For this situation we will consider two models, M_0 and M_1. Model M_0 is the linear regression model

$$\mu_i = \beta_0 + \beta_1(x_i - \bar{x}),$$

where $\bar{x} = n^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n_i} x_{ij}$ with $n = \sum_{i=1}^{c} n_i$. Model M_1 is a one-way analysis of variance model with

$$\mu_i = \beta_0 + \beta_i.$$

We collect the observations in a $n \times 1$ vector

$$\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{i1}, \ldots, Y_{in_i}, \ldots, Y_{c1}, \ldots, Y_{cn_c})^T,$$

and correspondingly write

$$\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{1n_1}, \ldots, \mu_{i1}, \ldots, \mu_{in_i}, \ldots, \mu_{c1}, \ldots, \mu_{cn_c})^T$$

for the $n \times 1$ vector of mean values. Further, the model matrices for models M_0 and M_1 are denoted $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ and the model spaces are denoted $C(\boldsymbol{X}_0)$ and $C(\boldsymbol{X}_1)$. We use the notation $\mathbf{1}_k$ to denote a $k \times 1$ vector of 1's.

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 $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ onto the two model spaces. (Here you may consider the results of section 2.3.2 in the textbook and additional exercise 3 as known.)

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

$$\widehat{\boldsymbol{\mu}}_0 = \bar{Y}\mathbf{1}_n + \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n),$$

and

$$\widehat{\boldsymbol{\mu}}_1 = (\bar{Y}_1, \ldots, \bar{Y}_1, \ldots, \bar{Y}_i, \ldots, \bar{Y}_i, \ldots, \bar{Y}_c, \ldots, \bar{Y}_c)^T,$$

where

$$\boldsymbol{x} = (x_1, \ldots, x_1, \ldots, x_i, \ldots, x_i, \ldots x_c, \ldots x_c)^T,$$

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij},$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{c} n_i\bar{Y}_i,$$

and

$$\widehat{\beta}_1 = M^{-1}\sum_{i=1}^{c} n_i\,(x_i - \bar{x})\,\bar{Y}_i$$

with $M = \sum_{i=1}^{c} n_i\,(x_i - \bar{x})^2$.

We now consider the orthogonal decomposition

$$\boldsymbol{Y} = [\boldsymbol{P}_0 + (\boldsymbol{P}_1 - \boldsymbol{P}_0) + (\boldsymbol{I} - \boldsymbol{P}_1)]\,\boldsymbol{Y}$$

and the corresponding sum-of-squares decomposition

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}.$$

d) Show that

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

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

$$F = \frac{\|\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0\|^2/(c-2)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_1\|^2/(n-c)} = \frac{\sum_{i=1}^{c} n_i[\bar{Y}_i - \bar{Y} - \widehat{\beta}_1(x_i - \bar{x})]^2/(c-2)}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n-c)}$$

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

We now return to the example on REM sleep and ethanol 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 a study it was recorded whether 35 patients experienced sore throat or not on waking after surgery. The aim of the study was to investigate how the risk of sore throat depends on the duration of the surgery and the type of device used to secure the airway during surgery.

The data file consists of one line for each of the 35 patients and with the following variables in the three columns:

- `duration`: duration of the surgery (in minutes)
- `type`: type of device used to secure the airway
  (0 = laryngeal mask airway, 1 = tracheal tube)
- `sore`: sore throat on waking after surgery (0 = no, 1 = yes)

We will use logistic regression to analyze 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/sore-throat.txt"
sore.throat=read.table(data,header=T)
```

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

b) Denote by β_1 the regression coefficient for `duration` in the logistic regression model. Show that $e^{10\beta_1}$ is the odds ratio corresponding to 10 minutes increase in the duration of the surgery. Estimate this odds ratio.

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

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) Extra and optional (does not count, as the relevant theory will be lectured the last week before the deadline): Find 95% confidence intervals for $e^{10\beta_1}$ from b) and for e^{β_2} from c). Give interpretations of these intervals.