STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 2 of 2

Submission deadline

Thursday November 6 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 R 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 look at data from a fertility study in Botswana, where the aim is to investigate how the number of living children a woman gives birth to depends on whether she has education, uses contraception, lives in an urban area, etc.

You may read the data into R by the commands:

```
data_file <- "http://www.uio.no/studier/emner/matnat/math/STK3100/data/fertility_data.csv"
fertility <- read.csv(data_file,header=TRUE)
```

The data file consists of 8 columns with the following variables:

- `educ0`: indicator of whether the woman has education (0 = no; 1 = yes)
- `usemeth`: indicator of whether the woman uses contraception (0 = no; 1 = yes)
- `urban`: indicator of whether the woman lives in an urban area (0 = no; 1 = yes)
- `electric`: indicator of whether the woman has electricity installed in her house (0 = no; 1 = yes)
- `radio`: indicator of whether the woman has a radio (0 = no; 1 = yes)
- `tv`: indicator of whether the woman has a TV (0 = no; 1 = yes)
- `bicycle`: indicator of whether the woman has a bicycle (0 = no; 1 = yes)
- `ceb`: number of living children.

In this problem, we will assume that the number of living children (`ceb`) is Poisson distributed within each of the $2^7 = 128$ combinations of the 7 binary covariates.

a) Expain why this may be a reasonable assumption.

b) Fit a GLM for Poisson data with logarithmic link function to the data, using all the covariates.

c) Perform an analysis that clarifies the significance of the different covariates, where you consider the effect of removing some of the covariates from the model. Which of the models you have considered seems to give the best description of the data?

d) Interpret the estimates from "the best model" in question c) as rate ratios, and give 95% confidence intervals for the rate ratios.

e) Estimate the claim rate of a woman who does not have education, does not use contraception, does not live in an urban area, has electricity, a radio and a bicycle, but no TV. Also give a 95% confidence interval for this rate.

Problem 2

The Poisson distribution has variance equal to the mean. In practice this assumption is often unrealistic for count data, because the variability is in fact greater than can be described by the Poisson mean. This is what we call *overdispersion*. A common way to handle overdispersed count data is to use a type of mixture of Poisson distributions, which results in the negative binomial distribution. In this problem we will consider some properties of the negative binomial distribution and the corresponding GLMs. As shown in the lectures, the negative binomial distribution may be obtained as a mixture of Poisson distributions.

More specifically, if Λ is a random variable that follows the gamma distribution with pdf

$$f(\gamma;\mu,k)=\frac{(k/\mu)^k}{\Gamma(k)}\lambda^{k-1}e^{-k\lambda/\mu},\quad \lambda>0$$

and further, the random variable Y, given $\Lambda=\lambda$, is Possion distributed with parameter λ, and thus has the conditional pmf

$$p(y|\lambda)=\frac{\lambda^y}{y!}e^{-\lambda},\quad y=0,1,2,\ldots$$

Then, the marginal pmf of Y is given by

$$p(y;\mu,k)=\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(\frac{\mu}{\mu+k}\right)^y,\quad y=0,1,2,\ldots, \qquad (1)$$

which is the pmf of the negative binomial distribution.

We will now assume that $k>0$ is a given constant, and consider the random variable $Y^*=Y/k$. Then $P(Y^*=y^*)=P(Y=ky^*)$, for $y^*=0,\frac{1}{k},\frac{2}{k},\ldots$, so Y^* has pmf

$$p(y^*;\mu,k)=\frac{\Gamma(ky^*+k)}{\Gamma(k)\Gamma(ky^*+1)}\left(\frac{k}{\mu+k}\right)^k\left(\frac{\mu}{\mu+k}\right)^{ky^*},\quad y=0,\frac{1}{k},\frac{2}{k},\ldots. \qquad (2)$$

a) Show that (2) is a distribution in the exponential dispersion family. That is, show that (2) can be written on the form $\exp((\theta y^*-b(\theta))/a(\phi)+c(y;\phi))$, with $a(\phi)=1/k$, and determine θ and $b(\theta)$.

b) Find the mean and variance of Y^* using the relations (4.3) and (4.4) in the text book. Use these results to show that $\mathrm{E}(Y)=\mu$ and determine $\mathrm{Var}(Y)$.

Then we assume that $Y_1, \ldots, Y_n$ are independent and have pmf of the form (1), and that their means $\mu_i = \mathrm{E}(Y_i)$ are specified via a link function g, i.e. $g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}$.

c) Derive an expression for the log-likelihood function $L(\boldsymbol{\mu}, k; \mathbf{y})$. (In the text book, there is an expression of the log-likelihood for the parameterisation with $\gamma = 1/k$. You should express it in terms of k.)

d) For a given $k > 0$, the deviance for a negative binomial GLM is given by $D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2(L(\mathbf{y}, k; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}, k; \mathbf{y}))$. Derive an expression for $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$.

e) Derive the limit of the deviance when $k \to \infty$. How can you explain this result?

We will now return to the fertility data from Problem 1, where it was assumed that the Poisson distribution was a good fit, and hence, that there was no overdispersion.

f) Fit your prefered GLM from Problem 1 c), substituting the Poisson distribution with the negative binomial (this is done using the command *glm.nb* from the `MASS` package, see the R code on the horseshoe crab data from the lecture on October 21). Does it provide a better fit than the Poisson GLM? What does the estimated k (called θ in the R output) tell you about possible over-dispersion, and how do you see that in light of your response to Problem 1 a)?