23rd October, 2023

STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 2 of 2

Submission deadline

Thursday 10th November 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.
- Give satisfactory answers in at least 60% of the (sub-)questions
- Include relevant R outputs in your report

Problem 1. In this problem we will look at data on the number of claims during one year in a portfolio of insured cars from an English insurance company. The number of claims are registered according to the age of the insured (given in four age groups), engine volume of the car (given in four volume groups), and the district where the car is insured (four districts).

You may read the data into R by the commands:

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

The data file consists of one line for each of the 64 combinations of age group, volume group and district, and with the following variables in the five columns:

- `alder`: age of policyholder (1 = below 25 years; 2 = 25–29 years; 3 = 30–35 years; 4 = over 35 years).
- `motorvolum`: Engine volume (1 = below 1 litre; 2 = 1–1.5 litres; 3 = 1.5–2 litres; 4 = over 2 litres).
- `distrikt`: District: (4 = London and other large cities; 1–3 = other districts).
- `antforsikret`: Number of insured cars.
- `antskader`: Number of claims.

We will assume that the number of claims is Poisson distributed within each of the 64 combinations of age group, volume group and district.

a) Expain why this may be a reasonable assumption.

We will use a GLM for Poisson data with logarithmic link function and age of the policyholder (`alder`), engine volume (`motorvolum`) and district (`distrikt`) as categorical covariates (factors).

b) Explain why you should use the logarithm of the number of insured cars (`antforsikret`) as an offset.

c) Perform an analysis that clarifies the significance of age, engine volume, and district and any potential interactions between these factors. Which of the models you have considered seems to give the best description of the data?

d) Make some informative plots of the residuals for "the best model" from question c. Are there any patterns in the plots which suggest that the model fit is not satisfactory?

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

f) Estimate the claim rate of an insured person in age category 25–29 years who has a car with engine volume 1.5–2 liter and lives in London. 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. We will start by showing how the negative binomial distribution may be obtained as a mixture of Poisson distributions.

There exists various formulations of the negative binomial pmf. In this problem we will assume that the pmf of the negative binomial distribution takes the form

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

To see how the negative binomial distribution may be obtained as a mixture of Poisson distributions, we assume that the random variable Λ is gamma distributed with pdf

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

and further that given $\Lambda = \lambda$, the random variable Y is Poisson distributed with parameter λ. Thus the conditional pmf of Y is given by

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

a) Show that (1) is the marginal pmf of Y.

We will 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

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

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

c) 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 $E(Y) = \mu$ and determine $\mathrm{var}(Y)$.

Then we assume that $Y_1, Y_2, \ldots, Y_n$ are independent and have pmf of the form (1), and that their means $\mu_i = E(Y_i)$ are specified via a link function g, i.e. $g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}$.

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

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

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