Cover Page

Department of Mathematical Sciences

Examination paper for TMA4315 - Generalized Linear Models

Examination date: 5th December 2023

Examination time (from-to): 15:00 - 19:00

Permitted examination support material: C
- Tabeller og formler i statistikk (Tapir forlag, Fagbokforlaget),
- one yellow A4 sheet with your own handwritten notes (stamped by the Department of Mathematical Sciences),
- specified calculator

Academic contact during examination: Bob O'Hara
Phone: 91554416

Academic contact present at the exam location: NO

OTHER INFORMATION

Get an overview of the question set before you start answering the questions.

Read the questions carefully and make your own assumptions. If a question is unclear/vague, make your own assumptions and specify them in your answer. The academic person is only contacted in case of errors or insufficiencies in the question set. Address an invigilator if you suspect errors or insufficiencies. Write down the question in advance.

Hand drawings: For question 12 you are meant to answer on handwritten sheets. Other questions should answered directly in Inspera. At the bottom of the question you will find a seven-digit code. Fill in this code in the top left corner of the sheets you wish to submit. We recommend that you do this during the exam. If you require access to the codes after the examination time ends, click “Show submission”.

Weighting: The maximum acheivable score for each question is given with the question.

Notifications: If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen.

Withdrawing from the exam: If you become ill or 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 in the archive in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.

A few years ago a group of researchers investigated the numbers of deaths caused by hurricanes, and concluded that hurricanes with female sounding names were deadlier.

Of course, one has to control for the strength of the hurricane, so a variable called NDAM (=“normalised damage”) was used.

We can assume that the number of deaths follows a Poisson distribution, and fit a GLM.

1 Parts of a GLM

List and explain the 3 parts of a Generalised Linear Model (1-2 sentences for each element)

Fill in your answer here

1. The random component, which is a distribution from the exponential family.
2. The link function, which transforms the systematic component to be the expected value of each variable in the random component (or the response function, which is the inverse of the link function). It should be invertible and twice differentiable.
3. The systematic component, which is a linear predictor, X^T \beta

---

Maximum marks: 6

The negative binomial distribution can be written in this form

$$f(y|r,p) = \binom{y+r-1}{y}(1-p)^r p^y$$

with y=0,1,2,..., r>0, 0<p<1. We are interested in modelling *p*, and we can treat *r* as a constant.

We should write this as a member of the exponential family:

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w_i + c(y, w_i, \phi)\right)$$

2 b()

What is $b(\theta)$?

Answer: - r log(1-p)

(or - r ln(1-p): at some point in maths they become equivalent). Also can write in terms of p=exp(theta)

Maximum marks: 2

3 Canonical Link Function

What is the canonical link function?
Select one alternative:

- inverse
- X log
- cloglog
- exponential
- probit
- logit
- identity
- something else

Maximum marks: 1

A few years ago a group of researchers investigated the numbers of deaths caused by hurricanes, and concluded that hurricanes with female sounding names were deadlier.
Of course, one has to control for the strength of the hurricane, so a variable called NDAM (="normalised damage") was used.
We can assume that the number of deaths follows a Poisson distribution, and fit a GLM.

When the model is fitted, we get the following output.

```
Call:
glm(formula = alldeaths ~ NDAM + Minpressure + Gender, family = "poisson",
    data = Himmicanes)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.88299    0.16187  17.810  < 2e-16 ***
NDAM         0.32597    0.07708   4.229 2.35e-05 ***
Minpressure -0.43588    0.14002  -3.113  0.00185 **
GenderMale  -0.58112    0.29224  -1.989  0.04675 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 28.03668)

    Null deviance: 4031.9  on 91  degrees of freedom
Residual deviance: 2467.2  on 88  degrees of freedom
AIC: 2800.1

Number of Fisher Scoring iterations: 6
```

NDAM is the amount of damage done by the hurricane (adjusted for inflation): a stronger hurricane should do more damage. It is standardised to mean 0, variance 1.
Minpressure is the minimum pressure of the hurricane: a stronger hurricane has a lower pressure. It is standardised to mean 0, variance 1.
Gender is the gender of the name assigned to the hurricane: it is a factor with two levels (*Male* and *Female*)

4 Gender Effect

If two hurricanes had the same NDAM and minpressure, by how many times would the prediced deaths be higher if the gender was female?

Answer to 2 decimal places: 1.79 .

Maximum marks: 1

5 Gender Effect Calculations

Show your working for this problem

Fill in your answer here

The male effect is -0.58, so the female effect is -(-0.58)

exp(-(-0.58))

Maximum marks: 3

A few years ago a group of researchers investigated the numbers of deaths caused by hurricanes, and concluded that hurricanes with female sounding names were deadlier.
Of course, one has to control for the strength of the hurricane, so a variable called NDAM (="normalised damage") was used.
We can assume that the number of deaths follows a Poisson distribution, and fit a GLM.

When the model is fitted, we get the following output.

```
Call:
glm(formula = alldeaths ~ NDAM + Minpressure + Gender, family = "poisson",
    data = Himmicanes)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.88299    0.16187  17.810  < 2e-16 ***
NDAM         0.32597    0.07708   4.229 2.35e-05 ***
Minpressure -0.43588    0.14002  -3.113  0.00185 **
GenderMale  -0.58112    0.29224  -1.989  0.04675 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 28.03668)

    Null deviance: 4031.9  on 91  degrees of freedom
Residual deviance: 2467.2  on 88  degrees of freedom
AIC: 2800.1

Number of Fisher Scoring iterations: 6
```

NDAM is the amount of damage done by the hurricane (adjusted for inflation): a stronger hurricane should do more damage. It is standardised to mean 0, variance 1.
Minpressure is the minimum pressure of the hurricane: a stronger hurricane has a lower pressure. It is standardised to mean 0, variance 1.
Gender is the gender of the name assigned to the hurricane: it is a factor with two levels (*Male* and *Female*)

6 Confidence Interval

Calculate the lower bound of a 95% confidence interval for this estimate.

Answer to 2 decimal places: 1.01 .

Maximum marks: 1

7 Confidence Interval Working

Show your working for this problem

Fill in your answer here

The 95% CI is exp(beta_0 +/-1.96*se)
=exp(0.58-1.96*0.29) = 1.01

Maximum marks: 3

A few years ago a group of researchers investigated the numbers of deaths caused by hurricanes, and concluded that hurricanes with female sounding names were deadlier.
Of course, one has to control for the strength of the hurricane, so a variable called NDAM (="normalised damage") was used.
We can assume that the number of deaths follows a Poisson distribution, and fit a GLM.

When the model is fitted, we get the following output.

```
Call:
glm(formula = alldeaths ~ NDAM + Minpressure + Gender, family = "poisson",
    data = Himmicanes)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.88299    0.16187  17.810  < 2e-16 ***
NDAM         0.32597    0.07708   4.229 2.35e-05 ***
Minpressure -0.43588    0.14002  -3.113  0.00185 **
GenderMale  -0.58112    0.29224  -1.989  0.04675 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 28.03668)

    Null deviance: 4031.9  on 91  degrees of freedom
Residual deviance: 2467.2  on 88  degrees of freedom
AIC: 2800.1

Number of Fisher Scoring iterations: 6
```

NDAM is the amount of damage done by the hurricane (adjusted for inflation): a stronger hurricane should do more damage. It is standardised to mean 0, variance 1.
Minpressure is the minimum pressure of the hurricane: a stronger hurricane has a lower pressure. It is standardised to mean 0, variance 1.
Gender is the gender of the name assigned to the hurricane: it is a factor with two levels (*Male* and *Female*)

8 Overdispersion

The output states ´(Dispersion parameter for poisson family taken to be 19.28544)´. The dispersion parameter would normally be around 1.

What does a value of 19.3 suggest about the data and the assumption that it follows a Poisson distribution? What might cause this?

Fill in your answer here

The value suggests that there is overdispersion, i.e. a lot more variation than from a Poisson distribution. This could be cause by all sorts of things, such as unobserved covariates, like the time of year or where the hurricane struck. There could also clustering in deaths, e.g. depending on whether the hurricane hit a large city like New Orleans.

Maximum marks: 4

9 Overdispersion Calculation

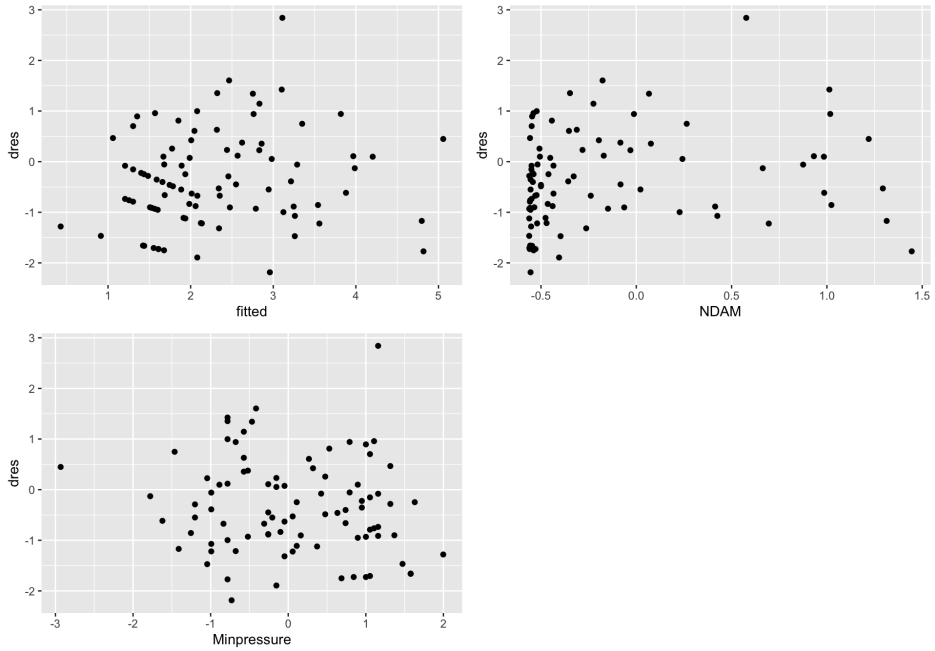
How could this value be calculated?

Fill in your answer here

residual deviance/residual df = 2467.2/88

Or from a chi-squared statistic: (\sum (O_i - E_i)^2/E_i)

Maximum marks: 2



The data analysis was re-examined and residual plots suggested there was a problem with the model. Here are the plots, of the deviance residuals against the fitted values and continuous covariates.

10 How good is the model?

Describe (briefly!) what these plots tells you about the model fit.

Fill in your answer here

This mostly looks OK: no obvious residuals (one point might be, but it's not clear). However, there seems to be a non-linear effect of NDAM: the residuals look curved in the plot.

Maximum marks: 4

11 Model Comparison

After looking at this, a better model was suggested, by adding terms into the model, so that the old model was nested within the new one. The two models were compared, giving the following output

```
Analysis of Deviance Table
Model 1: alldeaths ~ ...
Model 2: alldeaths ~ ....
  Resid. Df Resid. Dev Df Deviance
1        88     2467.2
2        86     2034.9  2   432.28
```

Is there evidence the new model is better? Explain which statistics you used to reach this conclusion.

Fill in your answer here

Yes, the new model is much better: the deviance is 432.28 on 2 degrees of freedom. Even without looking it up, this is VERY VERY significant.

(the new model is actually one with a quadratic term on NDAM)

Maximum marks: 4

12 LMMs

Note: Answers for this question must be written on handwritten sheets.

In a linear mixed model, the joint distribution of **Y** and $\boldsymbol{\gamma}$ is:

$$\begin{pmatrix}\mathbf{Y}\\ \boldsymbol{\gamma}\end{pmatrix} \sim N(\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{UGU}^T+\sigma^2\mathbf{I} & \mathbf{?}\\ \mathbf{?^T} & \mathbf{G}\end{pmatrix})$$

We want to look at the off-diagonal part, i.e. $Cov(\mathbf{Y}, \boldsymbol{\gamma})$, for a model with a random intercept, i.e.

$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \varepsilon_{ij}$
where $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ $\gamma_{0i} \sim N(0, \tau_0^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$

a. Show that when $i = k$, $Cov(y_{ij}, \gamma_{0k}) = E(\gamma_{0i}^2) = Var(\gamma_{0i})$.
b. Derive the expression for when $i \neq k$.
c. From this, explain what the matrix $\mathbf{?}$ looks like (or draw it if that is easier!)
We can look at this model:
$y_{ij} = x_{ij}\beta + \gamma_{0i} + \gamma_{1i}x_{ij} + \varepsilon_{ij}$ with $\gamma_{1i} \sim N(0, \tau_1^2)$ and $\mathrm{Corr}(\gamma_{0i}, \gamma_{1i}) = \rho$,
having a single covariate and a random slope for the effect of that covariate that varies between groups.
d. Derive $Cov(y_{ij}, y_{ik})$.

Maximum marks: 14

Biologists have collected data on the sizes of birds and the sizes of their eggs: an obvious expectation is that larger birds have larger eggs.

Species are grouped into genus (i.e. each genus contains more than one species), so we can look at whether the change between genera is greater than between species.

The following model was fitted

Egg1 <- lmer(logEggMass ~ logMassSpecies + logMassGenus + (1|Genus), data=Eggs)

where

- **logEggMass** is the log of the average egg mass (base e) for each species
- **logMassSpecies** is the log of the average body mass (base e) for each species, after correcting for genus mass (this correction is only important for teh final question)
- **logMassGenus** is the log of the average body mass (base e) for each genus
- **Genus** is the genus that a species belongs to

13 Equation

Write down the equation for this model (either in matrix or scalar form).

Y_i = \beta_0 + x_1i \beta_1 + \beta_2 x_2i + \gamma(j)

(if you struggle to use the maths tool, write it on paper)

Maximum marks: 3

14 Explain the equation

Explain how each part of the equation relates to the model in R code.

Fill in your answer here

Y_i = logEggMass (the response)
x_1i = logMassSpecies \beta_1 = coefficient for logMassSpecies
x_2j = logMassGenus \beta_2 = coefficient for logMassGenus j = genus
\gamma_j = (1|Genus) = genus random effect

Maximum marks: 3

When the model was fitted, the following summary was obtained

```
Linear mixed model fit by REML ['lmerMod']
Formula: logEggMass ~ MassSpecies + MassGenus + (1 | Genus)
   Data: Eggs

REML criterion at convergence: 98.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8334 -0.3899 -0.0264  0.3860  4.7210

Random effects:
 Groups   Name        Variance Std.Dev.
 Genus    (Intercept) 0.19338  0.4397
 Residual             0.02745  0.1657
Number of obs: 237, groups:  Genus, 101

Fixed effects:
            Estimate Std. Error t value
(Intercept) -1.17468    0.15623  -7.519
MassSpecies  0.54835    0.03261  16.817
MassGenus    0.75012    0.02484  30.193

Correlation of Fixed Effects:
            (Intr) MssSpc
MassSpecies  0.000
MassGenus   -0.956  0.000
```

15 Egg Size Effects

There is a theory that the effect of mass on egg size should be lower within a species than at the genus level, i.e. the coefficient for *MassSpecies* < *MassGenus*.

Test whether the data support this theory using the information provided in the model output. If you cannot do the test, explain how you could do it if you had time/the correct information.

Fill in your answer here (or write your answer on a piece of paper)

One way: use a Wald test to (1) calculate the difference (= constrast C), (2) extract the standard error (sqrt(V)). From that, calculate the test statistic (C' mu C/sqrt(V)), and compare it to the distribution under the null hypothesis.

```
Effs <-  unlist(coef(EggMS)$Genus[1,c("logFemaleMassSpecies", "logFemaleMassGenus")])
Contrast <- as.matrix(c(1,-1))
Var <- as.matrix(vcov(EggMS)[2:3,2:3])
Mu <- Contrast%*%Effs
Var <- t(Contrast)%*%Var%*%Contrast
t <- (Mu - 0)/sqrt(Var)
```

Maximum marks: 5

16 Likelihood Ratio Tests

It is possible to do a likelihood ratio test, with some work:

(a) how would you arrange the data to do the test, and what models would you fit?
(b) if you managed to obtain the correct models, how would you perform the test?

(you should be able to anser part b even if you cannot do part a)

Fill in your answer here

(a) The null hypothesis is that \beta_1 = \beta_2 = \beta_W (say). So the model is
Y_i = \beta_0 + x_1i \beta_W + \beta_W x_2i + \gamma(j) = \beta_0 + (x_1i + x_2i)*\beta_W + \gamma(j)
Thus, we add the Species and Genus weights and fit a model with the sum. We compare that to the full model (above), but we have to use 2\alpha as a critical value (because we are using a one-tailed test). Note we also have to check that the species effect is smaller than the genus effect.

(b) If we have constructed the models for the null and alternative hypotheses, M_0 and M_1,we can calculate the likelihoods for each of them, L_0 and L_1, and compare these. Nore that this also needs an estimate of the residual error, i.e. we use an F test.

Maximum marks: 8

Question 12 answer

Question 12

1. Show that when $i = k$, $\text{Cov}(y_{ij}, \gamma_{0k}) = \text{E}(\gamma_{0i}^2) = \text{Var}(\gamma_{0i})$.

First,

$$\text{Cov}(y_{ij}, \gamma_{0k}) = \text{E}(\mathbf{X}\boldsymbol{\beta} + \gamma_{0i} + \varepsilon_{ij} - \mathbf{X}\boldsymbol{\beta})\gamma_{0i}) = \text{E}(\gamma_{0i}^2) + \text{E}(\varepsilon_{ij}\gamma_{0i})$$

Because ε_{ij} & γ_{0i} are orthogonal, $\text{E}(\varepsilon_{ij}\gamma_{0i}) = 0$. Then $\text{Cov}(y_{ij}, \gamma_{0i}) = \text{E}(\gamma_{0i}^2) = \text{Var}(\gamma_{0k}) = \tau_0^2$ (note that $\text{E}(\gamma_{0k}) = 0$, so $\text{Var}(\gamma_{0k}) = \text{E}(\gamma_{0i}^2)$)

b. Derive the expression for when $i \neq k$.

In this case we have

$$\text{Cov}(y_{ij}, \gamma_{0k}) = \text{E}(\mathbf{X}\boldsymbol{\beta} + \gamma_{0k} + \varepsilon_{ij} - \text{E}(\mathbf{X}\boldsymbol{\beta}))\gamma_{0k}) = \text{E}(\gamma_{0i}\gamma_{0k}) + \text{E}(\varepsilon_{ij}\gamma_{0k})$$

And not only do we have $E(\varepsilon_{ij}\gamma_{0i}) = 0$, we also have $E(\gamma_{0i}\gamma_{0k}) = 0$, so this is 0.

c. From this, explain what the matrix ? looks like (or draw it if that is easier!)

It's a block diagonal matrix, with each block bneing a vector of length n_i

For the model $y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \gamma_{0i} + \gamma_{1i}\mathbf{x}_{ij} + \varepsilon_{ij}$

d. Derive $\text{Cov}(y_{ij}, y_{ik})$

$$\begin{aligned}
\text{Cov}(y_{ij}, y_{ik}) &= \text{E}(\mathbf{x}_{ij}\boldsymbol{\beta} + \gamma_{0i} + \gamma_{1i}\mathbf{x}_{ij} + \varepsilon_{ij} - \text{E}(\mathbf{x}_{ij}\boldsymbol{\beta}))(\mathbf{x}_{ik}\boldsymbol{\beta} + \gamma_{0i} + \gamma_{1i}\mathbf{x}_{ik} + \varepsilon_{ik} - \text{E}(\mathbf{x}_{ik}\boldsymbol{\beta})) \\
&= \text{E}(\gamma_{0i} + \gamma_{1i}\mathbf{x}_{ij} + \varepsilon_{ij})(\gamma_{0i} + \gamma_{1i}\mathbf{x}_{ik} + \varepsilon_{ik}) \\
&= \text{E}(\gamma_{0i}\gamma_{0i}) + \text{E}(\gamma_{1i}\mathbf{x}_{ij}\gamma_{1i}\mathbf{x}_{ik}) + \text{E}(\gamma_{oi}\gamma_{1i}\mathbf{x}_{ik}) + \text{E}(\gamma_{oi}\gamma_{1i}\mathbf{x}_{ij}) \\
&= \text{Var}(\gamma_{0i}) + \text{E}(\gamma_{1i}\mathbf{x}_{ij}\gamma_{1i}\mathbf{x}_{ik}) + \text{E}(\gamma_{oi}\gamma_{1i})(\mathbf{x}_{ik} + \mathbf{x}_{ij}) \\
&= \tau_0^2 + \tau_1^2\mathbf{x}_{ij}\mathbf{x}_{ik} + \sqrt{\tau_0^2\tau_1^2}\rho(\mathbf{x}_{ik} + \mathbf{x}_{ij})
\end{aligned}$$

(note that if $j = k$ we would also have an extra $+\sigma^2$: the question wasn't clear about that)