i

Front page

Department of Mathematical Sciences

Examination paper for **TMA4315 Generalized linear models**

Academic contact during examination: Jarle Tufto
Phone: 99 70 55 19

Examination date: Thursday, December 12, 2019
Examination time (from-to): 15:00 - 19:00
Permitted examination support material: C.

- *Tabeller og formler i statistikk* (Tapir forlag, Fagbokforlaget),
- one yellow A5 sheet with your own handwritten notes (stamped by the Department of Mathematical Sciences),
- specified calculator.

Other information:

- All answers must be justified, and relevant calculations provided.
- If needed, you're allowed to round the degrees of freedom up or down to nearest values tabulated in "Tabeller og formler i statistikk".
- The exam questions are only available in English since this is a course at master's level given in English.
- For each problem you may write your answer into Inspera, or use the provided paper sheets. Remember to correctly specify the code for each problem (available in lower right corner of the problem) on your paper sheet, so that the scanned sheets will be matched with the correct problem. It is advisable that you write down the code when starting to write on the sheet.

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

1

Problem 1a

Consider a random variable Y. In our course we have considered the univariate exponential family having distribution (probability density function for continuous variables and probability mass function for discrete variables)

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

where θ is called the *natural parameter* (or parameter of interest) and ϕ the *dispersion parameter*. In this problem we will let $w = 1$ and $\phi = 1$.

Suppose that Y is a random variable with probability mass function $f_Y(y) = (1-p)^{y-1}p$ for $y = 1, 2, \ldots$.

- Show that this distributions belongs to the exponential family and find its natural parameter θ and the function $b(\theta)$.
- What are the formulas connecting the mean and variance of Y to $b(\theta)$?
- Use these formulas to find the mean and variance of Y.

Fill in your answer here and/or use the paper sheets provided.

2 Problem 1b

Consider a generalized linear model (GLM) using the distribution in problem 1a for the response variable Y_i and the canonical choice of link function.

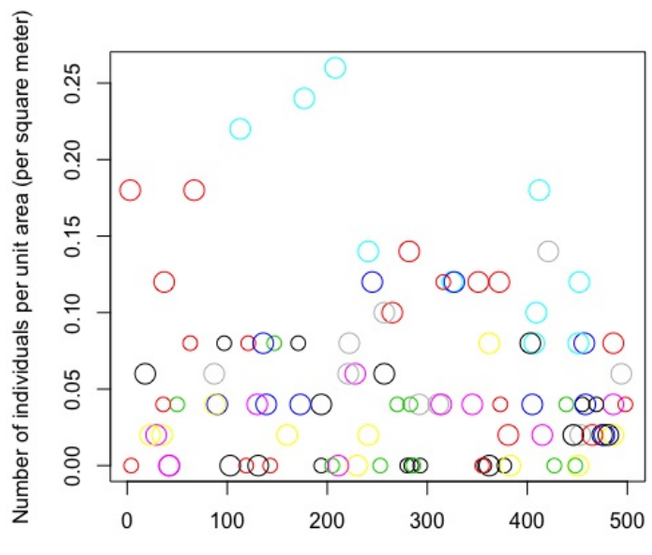
- Derive the score function $\boldsymbol{s}(\boldsymbol{\beta})$.
- Derive the expected Fisher information matrix $\boldsymbol{F}(\boldsymbol{\beta})$.
- Explain the idea behind Newton's method.
- Explain how the Fisher scoring algorithm differs from Newton's method and describe one potential advantage over Newton's method. Would these methods different for the glm considered here?
- Are there any constraints on the parameters $\beta_0, \beta_1, \ldots, \beta_k$ of the linear predictor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ in this particular glm? Would $\boldsymbol{\beta_0} = (0, 0, \ldots, 0)^T$ work as initial parameter values in the Fisher scoring algorithm?
- Explain how $\boldsymbol{F}(\boldsymbol{\beta})$ can be used to find the approximate standard errors of the maximum likelihood estimates.
  Fill in your answer here and/or use the paper sheets provided.

3 Problem 2a

A biologist studies the ecological niche of a rare plant species and wants to investigate if the species is more abundant at low altitudes. For each of a total number of 100 sampling locations, the number of plants within an area (of 25 or 50 square meters) was observed. The survey is taken between the years 2001 to 2010 but each location is observed only once. Also, the altitude (in meters above sea level) of each sampling location was recorded.

The first 20 observations are shown to the left below and a plot of number of number of individuals per square meter at each sampling location against altitude is shown in the plot to the right with different colours representing different years.

```
> head(data,20)
    y altitude year area
1   0      293 2001   25
2   0        4 2002   25
3   2      147 2003   25
4   2      139 2004   50
5   4      407 2005   50
6   2      130 2006   50
7   4      362 2007   50
8   1      453 2008   50
9   1      475 2009   50
10  6       37 2010   50
11  0      377 2001   25
12  0      143 2002   25
13  1       50 2003   25
14  1      477 2004   50
15 13      208 2005   50
16  3      228 2006   50
17  1      486 2007   50
18  2      292 2008   50
19  1      481 2009   50
20  1      381 2010   50
```



The variable **`y`** is the number of individuals of the species inside the different sampling squares, **`altitude`** is the meters above sea level of the sampling site, **`year`** is the year of sampling and **`area`** is the area of the sampling square measured in square meters.

We fit the following generalized linear model to the data.

```
> mod <- glm(y ~ altitude, offset=log(area), family=poisson, data=data)
> summary(mod)

Call:
glm(formula = y ~ altitude, family = poisson, data = data, offset = log(area))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6134  -1.5539  -0.5521   0.6644   4.1813

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.6592959  0.1259084 -21.121   <2e-16 ***
altitude    -0.0005859  0.0004150  -1.412    0.158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 237.42  on 99  degrees of freedom
Residual deviance: 235.43  on 98  degrees of freedom
AIC: 454.18

Number of Fisher Scoring iterations: 5

> logLik(mod)
'log Lik.' -225.0899 (df=2)
```

- Describe the model in appropriate mathematical notation and state its assumptions.
- Give an interpretation of estimated coefficient for **`altitude`** in the way you would communicate this estimate to a non-statistician.

Fill in your answer here and/or use the paper sheets provided.

4

Problem 2b

We have assumed that the number of individuals within each sampling square follows a Poisson distribution.

- What kind of underlying process would generate this distribution?
- Why is it reasonable to include **`log(area)`** as an offset variable in the model rather than as an ordinary covariate?

Fill in your answer here

5

Problem 2c

- Test if there is overdispersion in the data based on the fitted model in 2a.
- Briefly discuss possible biological mechanisms that could generate over-dispersion in the present data set.
- Estimate the overdispersion parameter from the observed deviance and use quasi-likelihood theory to test the significance of the effect of altitude using a level of significance equal to 0.05.

Fill in your answer here and/or use the paper sheets provided.

6 Problem 2d

To account for variation between years we fit a generalized linear mixed model with a random effect on the intercept using the variable **`year`** as a grouping factor as follows:

```
> mod2 <- glmmTMB::glmmTMB(y ~ altitude + (1|year), offset=log(area),
+                          family=poisson, data=data)
> summary(mod2)
 Family: poisson  ( log )
Formula:          y ~ altitude + (1 | year)
Data: data
 Offset: log(area)

     AIC      BIC   logLik deviance df.resid
   361.1    368.9   -177.5    355.1       97

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 year   (Intercept) 0.4405   0.6637
Number of obs: 100, groups:  year, 10

Conditional model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.7793729  0.2509643 -11.075  < 2e-16 ***
altitude    -0.0012530  0.0004379  -2.861  0.00422 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- State the assumptions of the model in appropriate mathematical notation.
- What is the maximum likelihood estimate of the variance of the random year effect?
- Conditional on the random year effect being zero, what is the expected number of individuals in a sampling site located at sea level and having an area of 50 square meters?
- Derive and evaluate an expression for the expected value of the same quantity (the number of individuals in a sampling site locate at sea level and having an area of 50 square meters) in a randomly chosen year. Hint: You'll need to make use of the law of total expectation and known properties of the lognormal distributions (see "Tabeller og formler i statistikk").

Fill in your answer here and/or use the paper sheets provided.

7 Problem 2e

We want to test if including the random effect in the above GLMM significantly improves the model.

- Formulate the appropriate null and alternative hypothesis in terms of the unknown model parameter of interest.
- What is the expected value of the component of the score vector associated with this parameter when we evaluate the score vector at the true parameter values ?
- If the null hypothesis is true, what is the approximate/asymptotic distribution of twice the difference between the maximum log likelihoods under the null and alternative hypothesis?
- Find the rejection region of the above test for a level of significance equal to 0.005 (see "Tabeller og formler i statistikk").
- Compute the observed value of the above likelihood ratio test statistic using output given in previous questions.
- What is the conclusion of the test?

Fill in your answer here

8

Problem 2f (multiple choice)

Suppose that we instead of the number of individuals y_i inside each sampling square only observed the binary response variable

$$z_i = \begin{cases} 0 \text{ for } y_i = 0 \\ 1 \text{ for } y_i \geq 1. \end{cases}$$

We keep the other model assumptions are unchanged. To estimate the same parameters as in 2a, what type of statistical model would you need to use?

Select one alternative:

- A GLM with a binary response and a cloglog link function, log(area) as an offset and altitude as covariate.
- A GLM with a probit link function and area as an offset and altitude as the covariate
- A GLMM with a binary random effect.
- A LM with $\ln z_i$ as the response.

9

Problem 3a

Consider a generalized linear mixed model (GLMM) with a single grouping factor. Let $f(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}_i)$ denote the the conditional joint probability mass function of the observed responses in the *i*th cluster. The random effects $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_m$ are identically independently distributed across the m clusters, each having probability density function $f(\boldsymbol{\gamma}_i|\boldsymbol{\theta})$. The vector $\boldsymbol{\theta}$ contains the variance parameters specifying the distribution of the random effects and the vector $\boldsymbol{\beta}$ are the fixed effect regression coefficients.

- Derive an expression for the likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\theta})$ of the model in terms of $f(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}_i)$ and $f(\boldsymbol{\gamma}_i|\boldsymbol{\theta})$.
- Briefly describe in words two ways in which approximations of this likelihood can be computed in practice.
- Give the definition of the restricted (REML) likelihood function for a GLMM in terms of $L(\boldsymbol{\beta}, \boldsymbol{\theta})$. What is the potential advantage of the REML likelihood over the ordinary likelihood for a GLMM?

Fill in your answer here and/or use the paper sheets provided.

10

Problem 3b

Consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is a $n \times p$ matrix ($p < n$) with full rank and $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. For such models, the restricted likelihood (REML) can be defined as the likelihood of the $n - p$ error contrasts contained in the vector $\mathbf{w} = \mathbf{A}^T\mathbf{y}$. Here, $\mathbf{A}$ is a $n \times (n - p)$ matrix satisfying $\mathbf{A}\mathbf{A}^T = \mathbf{I}_n - \mathbf{H}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{n-p}$ (that is, $\mathbf{A}$ has orthogonal columns), and $\mathbf{A}^T\mathbf{X} = \mathbf{0}$ (that is, the columns of $\mathbf{A}$ and $\mathbf{X}$ are orthogonal).

- Show that $\mathbf{w}$ has a distribution that don't depend on $\boldsymbol{\beta}$.
- Derive the REML likelihood.
- Derive the REML estimator of σ^2.

Fill in your answer here and/or use the paper sheets provided.