NTNU
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4315 Generalized linear models**

Academic contact during examination: Mette Langaas

Phone: 988 47 649

Examination date: 7 December 2017

Examination time (from–to): 09:00–13:00

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.
The exam questions are only available in English since this is a course at master's level given in English.
For each problem the maximum possible score is noted.

Language: English

Number of pages: 6

Number of pages enclosed: 0

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Date Signature

Problem 1 Poisson regression

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*.

The Poisson distribution is a discrete distribution with probability mass function

$$f(y) = \frac{\lambda^y}{y!} \exp(-\lambda), \text{ for } y = 0, 1, \ldots$$

where $\lambda > 0$.

a) [10 points]

Show that the Poisson distribution is a univariate exponential family, and specify what the elements of the exponential family (θ, ϕ, $b(\theta)$, w, $c(y, \phi, w)$) are.
What is the connections between $\mathrm{E}(Y)$ and elements of the exponential family?
What is the connections between $\mathrm{Var}(Y)$ and elements of the exponential family?
Use these connections to derive the mean and variance for the Poisson distribution.
If the Poisson distribution is used as the distribution for the response in a generalized linear model, what is the *canonical link* function?

We consider a Poisson regression with log link $\eta_i = g(\mu_i) = \ln(\mu_i)$, and linear predictor equal to $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Further, let p be the number of regression parameters in $\boldsymbol{\beta}$ (intercept included). The response–covariate pairs $(Y_i, \boldsymbol{x}_i)$ are independent for $i = 1, \ldots, n$.

b) [15 points]

Does this set-up satisfy the *requirements* of a GLM model? Explain.
Write down the log-likelihood.
From the log-likelihood, *derive* the formula for the score function $\boldsymbol{s}(\boldsymbol{\beta})$ and the expected Fisher information matrix, $\boldsymbol{F}(\boldsymbol{\beta})$. What are the dimensions of $\boldsymbol{s}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$?
How can $\boldsymbol{s}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$ be used to arrive at a maximum likelihood estimate for $\boldsymbol{\beta}$?

We now look at a data set giving the number of species of tortoise on the various Galapagos Islands[1].

The data set contains measurements on 30 islands, and we study the following variables:

- `Species`: The number of species of tortoise found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `Nearest`: The distance from the nearest island (km).
- `Scruz`: The distance from Santa Cruz island (km).
- `Adjacent`: The area of the adjacent island (km^2).

We have fitted a Poisson regression with log link to `Species` as response, and the other five variables are used as continuous covariates. Print-out from the fitted model is given in Figure 1.

Let $\boldsymbol{\beta}$ be a 6×1 column vector with the regression coefficients (intercept included), and let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for $\boldsymbol{\beta}$.

c) [10 points]

Write down the asymptotic distribution for $\hat{\boldsymbol{\beta}}$, and specify how the covariance matrix for $\hat{\boldsymbol{\beta}}$ is estimated.

We will focus on the effect of `Elevation`, and denote the corresponding regression coefficient β_2. Write down the maximum likelihood estimate for β_2 from Figure 1.
How can you explain this value to a biologist interested in understanding the effect of `Elevation` on the number of species of tortoise found on the islands?
What is numerical value for the estimated standard deviation of $\hat{\beta}_2$ given in Figure 1?
Construct an approximate 95% confidence interval for β_2.

[1]Data taken from the book "Practical Regression and Anova using R" by Julian J. Faraway.

```
Call:
glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
    Adjacent, family = poisson(link = log), data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

Figure 1: Print-out from `summary` for fitting a Poisson regression to the Galapagos dataset.

Problem 2 Binary regression with grouped data

We will study a logistic regression on data from the Titanic ship accident[2], with `Survived` or `NotSurvived` as our binary response measured for each person onboard. Three categorical covariates are studied (coding in parentheses):

- `Class` (for the person's cabin on the ship): `Class1`, `Class2`, `Class3` and `Crew` (effect coding, also called sum-zero contrast)
- `Sex`: `Male`, `Female` (dummy coding with `Male` as reference)
- `Age`: `Child`, `Adult` (dummy coding with `Child` as reference)

The full data set is listed in the upper part of Figure 2. There were 14 combinations for `Class`, `Sex` and `Age` observed (the combination `Child` and `Crew` did not exist). A logistic regression was fitted to the data and print-out is in the lower part of Figure 2.

a) [15 points]

Write down the model assumptions for this logistic regression.
What is the estimated coefficient for the level `Crew` for the covariate `Class`?
Explain how you would proceed to test whether the covariate `Class` is significant. (You do not have sufficient information to perform this test.)
How would you explain what the effect of `AgeAdult` is in this model?
Perform a deviance test, at significance level 0.05, to assess the model fit. What is your conclusion?

[2]The data were taken from the table "Titanic" from "datasets" packages at CRAN.

```
> print(dss)
Survived NotSurvived Class    Sex   Age
       5           0   1st   Male Child
      11           0   2nd   Male Child
      13          35   3rd   Male Child
       1           0   1st Female Child
      13           0   2nd Female Child
      14          17   3rd Female Child
      57         118   1st   Male Adult
      14         154   2nd   Male Adult
      75         387   3rd   Male Adult
     192         670  Crew   Male Adult
     140           4   1st Female Adult
      80          13   2nd Female Adult
      76          89   3rd Female Adult
      20           3  Crew Female Adult
> fit = glm(cbind(Survived,NotSurvived) ~ Class+Sex+Age,
+   family = binomial(link = logit),
+   data = dss,contrasts = list("Class"=contr.sum))
> summary(fit)
glm(formula = cbind(Survived, NotSurvived) ~ Class + Sex + Age,
family = binomial(link = logit), data = dss,
contrasts = list(Class = contr.sum))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1356  -1.7126   0.7812   2.6800   4.3833

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.22806    0.23706  -0.962    0.336
Class1       0.91338    0.11024   8.286  < 2e-16 ***
Class2      -0.10471    0.11803  -0.887    0.375
Class3      -0.86438    0.09826  -8.797  < 2e-16 ***
SexFemale    2.42006    0.14041  17.236  < 2e-16 ***
AgeAdult    -1.06154    0.24403  -4.350 1.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 671.96  on 13  degrees of freedom
Residual deviance: 112.57  on  8  degrees of freedom
AIC: 171.19
Number of Fisher Scoring iterations: 5
```

Figure 2: Print-out from Titanic data and `summary` for fitting a logistic regression.

Problem 3 Random intercept linear mixed effects model

New on the reading list this year is how to handle correlated data in a regression setting, with the aid of the linear mixed effects model. The simplest version of such a model is the *random intercept model*.

a) [20 points]

Write a short introduction to the random intercept linear mixed effects model and its practical usage, for a student with a good background in multiple linear regression. The introduction should include an example and emphasis should be on:

- Model assumptions.
- The conditional and marginal model.
- What is the intraclass correlation and how can it be calculated from a given model fit?
- How are the regression coefficients estimated?

Topics you do *not* need to address are: REML, hypothesis testing, AIC.