Front page

Department of Mathematical Sciences

Examination paper for **TMA4315 Generalized linear models**

Academic contact during examination: Jarle Tufto
Phone: 99 70 55 19
Zoom: https://NTNU.zoom.us/j/94042781252?pwd=YmZEVHpPR3FrNElIQ0lnV0hlbEpRZz09

Technical support during examination: Orakel support services
Phone: 73 59 16 00

Examination date: Tuesday, December 15, 2020
Examination time (from-to): 09:00 - 13:00 (+ half an hour of extra time for scanning)

Permitted examination support material: A.

- All printed and hand-written support material including material on the Internet is allowed. All calculators and computers allowed. All forms of communication and collaboration with other students or anyone else is not allowed.

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" or you can use R.
- 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 paper sheets, instructions below.

Make your own assumptions: If a question is unclear/vague, make your own assumptions and specify them in your answer. Only contact academic contact in case of errors or insufficiencies in the question set.

Saving: Answers written in Inspera Assessment are automatically saved every 15 seconds. If you are working in another program remember to save your answer regularly.

Cheating/Plagiarism: The exam is an individual, independent work. Examination aids are permitted. During the exam it is not permitted to communicate with others about the exam questions, or distribute drafts for solutions. Such communication is regarded as cheating. All submitted answers will be subject to plagiarism control. *Read more about cheating and plagiarism here*.

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. All candidates will also receive an SMS to ensure that nobody misses out on important information. Please keep your phone available during the exam.

ABOUT SUBMISSION

File upload: The file containing answers to the questions must be uploaded before the examination time expires. 30 minutes are added to the examination time to manage the sketches/calculations/files. (The additional time is included in the remaining examination time shown in the top left-hand corner.)

How to digitize your sketches/calculations

How to create PDF documents *(probably not relevant)*

Remove personal information from the file(s) you want to upload *(probably not relevant)*

NB! You are responsible to ensure that you upload the correct file. Check the file you have uploaded by clicking “Download”. The file can be removed or replaced as long as the test is open.

The additional 30 minutes are reserved for submission. If you experience technical problems during upload/submission, you must contact technical support before the examination time expires. If you can’t get through immediately, hold the line until you get an answer.

Your answer will be submitted automatically when the examination time expires and the test closes, if you have answered at least one question. This will happen even if you do not click “Submit and return to dashboard” on the last page of the question set. You can reopen and edit your answer as long as the test is open. If no questions are answered by the time the examination time expires, your answer will not be submitted.

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.

Accessing your answer post-submission: You will find your answer in Archive when the examination time has expired.

1

Scanned pdf upload

When you have finished answering all the problems, scan your handwritten text and upload the pdf file using the link below. You may alternatively write directly into the forms below the questions on the following pages.

2

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 exponentially distributed with probability density function $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

- 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.
- The cumulant generating function of a random variable Y belonging to the exponential family can be written as $K_Y(u) = \frac{w}{\phi}\left(b(\theta + \frac{\phi}{w}u) - b(\theta)\right)$. Use this to find the third central moment $E((Y - EY)^3)$ when *Y* is exponentially distributed.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

3 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.

- Which function is the canonical link?
- Derive the score function $s(\boldsymbol{\beta})$.
- Derive the expected Fisher information matrix $F(\boldsymbol{\beta})$.
- 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?
- Give an alternative choice of link function that does not lead to any constraints on $\beta_0, \beta_1, \ldots, \beta_k$.
- **Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.**

4 Problem 2a

In this problem we will consider an ordinal regression model known as the proportional odds model to analyse letter grades at the ordinary and the re-sit ("continuation") exam in TMA4100 Mathematics I at NTNU from the years 2004 to 2008. The following R-session output shows the data and the fitted model. Note that $Y_i = 1$ denotes the grade A, $Y_i = 2$ denotes a B and so on.

```
> grades
   year cont   A   B   C   D   E   F
1  2004   no  78 166 322 289 244 249
2  2005   no  89 129 229 268 288 311
3  2006   no 153 172 345 172 124 393
4  2007   no  79 203 395 248 203 293
5  2008   no 113 264 465 277 138 305
6  2004  yes   0   1   8  26  68 113
7  2005  yes   0   1   2  10  43  68
8  2006  yes   0   3   5  36  51  59
9  2007  yes   0   3  15  19  53 115
10 2008  yes   1   1  21  26  40  84
```

```
> summary(mod)

Call:
vglm(formula = cbind(A, B, C, D, E, F) ~ year + cont, family = cumulative(parallel = TRUE,
    link = "logit"), data = grades)

Pearson residuals:
                      Min      1Q  Median      3Q   Max
logitlink(P[Y<=1]) -2.604 -1.3542 -1.1881 -0.78012 5.476
logitlink(P[Y<=2]) -2.593 -1.3799 -0.6753 -0.05455 1.173
logitlink(P[Y<=3]) -4.463 -3.1871 -1.6116  1.49573 3.601
logitlink(P[Y<=4]) -3.425 -2.6229 -0.6032  0.53423 3.572
logitlink(P[Y<=5]) -9.982 -0.2276  2.4292  3.83860 5.496

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.64050    0.06169 -42.803  < 2e-16 ***
(Intercept):2 -1.44225    0.05053 -28.543  < 2e-16 ***
(Intercept):3 -0.25275    0.04725  -5.350 8.81e-08 ***
(Intercept):4  0.50802    0.04753  10.687  < 2e-16 ***
(Intercept):5  1.28834    0.04971  25.917  < 2e-16 ***
year2005      -0.26024    0.06542  -3.978 6.95e-05 ***
year2006       0.11369    0.06430   1.768    0.077 .
year2007       0.07321    0.06320   1.158    0.247
year2008       0.31678    0.06219   5.094 3.50e-07 ***
cont yes      -1.58531    0.06964 -22.763  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  5

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
logitlink(P[Y<=3]), logitlink(P[Y<=4]), logitlink(P[Y<=5])

Residual deviance: 424.2567 on 40 degrees of freedom

Log-likelihood: -338.9404 on 40 degrees of freedom

Number of Fisher scoring iterations: 5
```

- State the assumptions of the fitted model in appropriate mathematical notation.
- What are the numerical estimates of the different model parameters?

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

5

Problem 2b

- When we compare the model predictions for the ordinary exams in the the year 2005 relative to 2004, how does the odds of passing the exam change, that is, the odds of the event $Y_i \leq 5$? Similarly, how does the odds of getting a grade of B or better change?
- Based on the fitted model in point a), by how many % does the odds of passing the exam decrease when we compare re-sit ("continuation") exams relative to ordinary exams? Does the decrease in percentage depend on which year we are considering?
- Compute an estimate of the probability of getting the grade B at the ordinary exam in the year of 2008. As a sanity check, compare this estimate to the observed count.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

6

Problem 2c

- Explain the meaning of the saturated model for the data considered in point a). How many parameters does it involve?
- Given the estimated probabilities $\hat{\pi}_{ir}$ for the model fitted in point a) and the observed counts y_{ir}, derive an expression for the deviance of the model.
- What is the distribution and expected value of the model deviance under the null hypothesis that the model is correct?
- Test the goodness-of-fit of the fitted model using a level of significance of $\alpha = 0.05$.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

7

Problem 2d

Next we will consider the following model including an interaction term in addition to the main effects of `year` and `cont`.

```
> summary(mod1)

Call:
vglm(formula = cbind(A, B, C, D, E, F) ~ year + cont + year:cont,
    family = cumulative(parallel = TRUE, link = "logit"), data = grades)

Pearson residuals:
                      Min      1Q  Median      3Q   Max
logitlink(P[Y<=1]) -2.660 -1.3569 -1.1964 -0.7434 5.789
logitlink(P[Y<=2]) -2.542 -1.3372 -0.7320 -0.1771 1.466
logitlink(P[Y<=3]) -4.830 -3.1469 -1.4516  1.3718 3.986
logitlink(P[Y<=4]) -3.606 -2.1777 -0.4368  0.1996 3.494
logitlink(P[Y<=5]) -9.555 -0.2926  2.8853  3.8226 4.261

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept):1     -2.63364    0.06336 -41.564  < 2e-16 ***
(Intercept):2     -1.43558    0.05255 -27.320  < 2e-16 ***
(Intercept):3     -0.24611    0.04939  -4.983 6.25e-07 ***
(Intercept):4      0.51488    0.04967  10.366  < 2e-16 ***
(Intercept):5      1.29637    0.05178  25.036  < 2e-16 ***
year2005          -0.27466    0.06858  -4.005 6.21e-05 ***
year2006           0.06950    0.06787   1.024   0.3059
year2007           0.08772    0.06713   1.307   0.1913
year2008           0.31884    0.06572   4.852 1.22e-06 ***
cont yes          -1.64040    0.14146 -11.596  < 2e-16 ***
year2005:cont yes  0.14797    0.23243   0.637   0.5244
year2006:cont yes  0.36685    0.21042   1.743   0.0813 .
year2007:cont yes -0.13297    0.20188  -0.659   0.5101
year2008:cont yes -0.03874    0.20537  -0.189   0.8504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  5

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
logitlink(P[Y<=3]), logitlink(P[Y<=4]), logitlink(P[Y<=5])

Residual deviance: 417.0221 on 36 degrees of freedom

Log-likelihood: -335.3231 on 36 degrees of freedom

Number of Fisher scoring iterations: 6
```

- Give a brief verbal explanation of the meaning of the interaction term.
- Based on this model, compute an estimate of how many % the odds of passing decrease when we compare a re-sit ("continuation") exam relative to an ordinary exam for the year 2005.
- Test if the interaction term is significant at the $\alpha = 0.05$ level of significance.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

8

Problem 3a

In this problem we will analyse the birthweight of in total 322 rat pups (newborn rats) coming from 27 different litters. Below are the first 20 observations showing the birth weights of 12 pups from the first litter and the first 8 birth weights from the second litter. Also show is the sex of each pup encoded as 0 for males and 1 for females.

```
> head(data,20)
   weight sex Litter Lsize Treatment
1    6.60   0      1    12   Control
2    7.40   0      1    12   Control
3    7.15   0      1    12   Control
4    7.24   0      1    12   Control
5    7.10   0      1    12   Control
6    6.04   0      1    12   Control
7    6.98   0      1    12   Control
8    7.05   0      1    12   Control
9    6.95   1      1    12   Control
10   6.29   1      1    12   Control
11   6.77   1      1    12   Control
12   6.57   1      1    12   Control
13   6.37   0      2    14   Control
14   6.37   0      2    14   Control
15   6.90   0      2    14   Control
16   6.34   0      2    14   Control
17   6.50   0      2    14   Control
18   6.10   0      2    14   Control
19   6.44   0      2    14   Control
20   6.94   0      2    14   Control
```

To analyse the data we fit an linear mixed model including a random intercept based on litter as grouping variable to model random variation in expected weights between litters and also a random slope for sex, also based on litter as the grouping variable, to model potential variation between litters in the difference in weight between the sexes.

```
> mod <- lmer(weight ~ 1 + sex + (1 + sex|Litter), data=data, REML=TRUE)
boundary (singular) fit: see ?isSingular
> summary(mod)
Linear mixed model fit by REML ['lmerMod']
Formula: weight ~ 1 + sex + (1 + sex | Litter)
   Data: data

REML criterion at convergence: 414.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.3283 -0.4794 -0.0062  0.5755  2.8736

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Litter   (Intercept) 0.42408  0.6512
          sex         0.02152  0.1467   -1.00
 Residual             0.15999  0.4000
Number of obs: 322, groups:  Litter, 27

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.40150    0.13032  49.123
sex         -0.39384    0.05546  -7.102

Correlation of Fixed Effects:
    (Intr)
sex -0.653
convergence code: 0
boundary (singular) fit: see ?isSingular
```

- State the precise assumptions of the model in suitable mathematical notation.
- Identify the estimates of all model parameters in the model summary above.
- What is REML estimate of the covariance between the random intercept and slope?

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

9

Problem 3b

The following gives the maximum ordinary and restricted log likelihoods for a number of model alternatives.

```
> logLik(lmer(weight ~ 1 + sex + (1 + sex|Litter), data=data, REML=TRUE))
boundary (singular) fit: see ?isSingular
'log Lik.' -207.1693 (df=6)
> logLik(lmer(weight ~ 1 + sex + (1|Litter), data=data, REML=TRUE))
'log Lik.' -210.2051 (df=4)
> logLik(lmer(weight ~ 1 + (1|Litter), data=data, REML=TRUE))
'log Lik.' -234.3691 (df=3)
>
> logLik(lmer(weight ~ 1 + sex + (1 + sex|Litter), data=data, REML=FALSE))
boundary (singular) fit: see ?isSingular
'log Lik.' -203.7879 (df=6)
> logLik(lmer(weight ~ 1 + sex + (1|Litter), data=data, REML=FALSE))
'log Lik.' -206.8217 (df=4)
> logLik(lmer(weight ~ 1 + (1|Litter), data=data, REML=FALSE))
'log Lik.' -233.0623 (df=3)
> logLik(lm(weight ~ 1 + sex, data=data))
'log Lik.' -309.5149 (df=3)
```

- Using the above model alternatives, test if there is a significant difference between the average birth weight of male and female pups, using a level of significance of $\alpha = 0.05$.
- Test if the random intercept term is significant (assuming that there is no random slope in the model). Base the test on the asymptotic distribution of a likelihood ratio statistic. Compute the critical value of the test.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

10

Problem 3c

Given that we decide to include the random intercept term in the model, we may want to test if the term for the random slope is statistically significant.

- State the null and alternative hypothesis of this test. What are the additional unknown parameters estimated under the alternative hypothesis?
- Are there any constraints that the additional parameters under the alternative hypothesis need to satisfy?
- What is the asymptotic distribution of the likelihood ratio statistic? Here you only need to refer to known theory for this.
- Based on the observed maximum likelihoods in point b), carry out a test. Compute the p-value of the test, and conclude the test based on the p-value, again using the a level of significance of $\alpha = 0.05$.

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.

11

Problem 3d

Consider a male and female pup from from a randomly chosen litter.

- Based on the model considered in point a), derive an expression for the covariance between their weights.
- Compute an estimate of this covariance based on the parameter estimates in point a).

Fill in your answer here or in the handwritten document to be scanned and uploaded before the end of the exam.