Cover Page

Examination paper for TMA4315 - Generalized Linear Models

Course contact: Bob O'Hara
Present at the exam location: NO

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

OTHER INFORMATION

Read the questions carefully and make your own assumptions. Specify in your answer which assumptions you have used as a basis for interpreting/defining the assignment.

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.

SPECIFIC INFORMATION FOR YOUR COURSE

Hand drawings:

For question 9 you are meant to answer on handwritten sheets. Other questions must be 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”.

You are responsible for filling in the correct codes on the handwritten sheets. Therefore, read the cover sheet carefully. The Examination Office cannot guarantee that that incorrectly completed sheets will be added to your assignment.

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

Notifications:

Any messages during the exam (e.g., in case of errors in the exam set) will be sent out via notifications in Inspera. A notification will appear as a dialog box on the screen. You can find the notification again by clicking on the bell icon at the top right

Withdrawing from the exam:

If you 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.

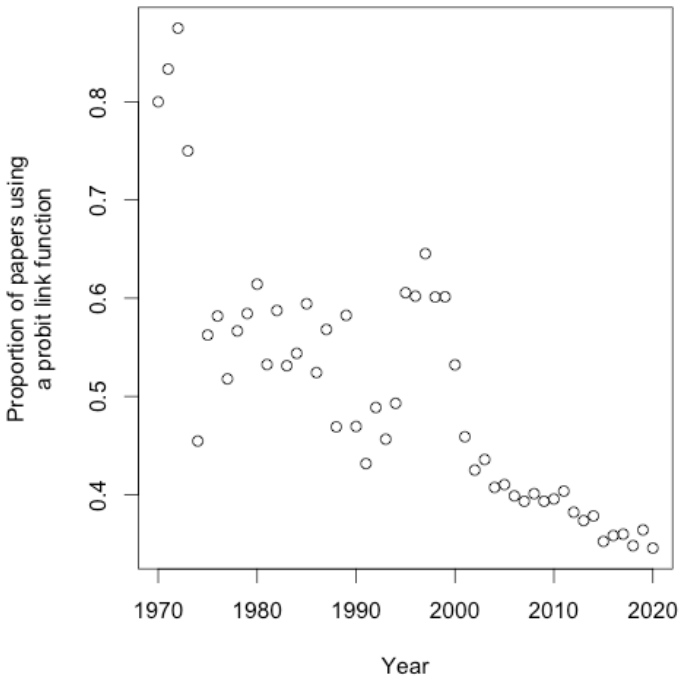
Access to your answers:

After the exam, you can find your answers under previous tests in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.

1 Copy of 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

Maximum marks: 6



Three link function are commonly used for the binomial distribution: the logit, probit and cloglog. Here we will compare logits and probits.

We can look at how the use of the logit and probit links has changed over time. The plot is of the proportion of papers mentioning 'probit' out of all papers mentioning 'probit' or 'logit' in each year.

2 What is a logit?

If η is a linear predictor, what is the response function for the logit link?

Maximum marks: 1

3 Describe the Data

Describe the general features of the data. Does it look like a model with a linear effect of Year (on the link scale) will fit well?

Fill in your answer here

Maximum marks: 3

Two models were fitted, one with a logit link and another with a probit link. These are the output from the summary() calls.

Note that YearC = Year - 1970, so 1970 is Year 0.

1. logit

```
Call:
glm(formula = cbind(Data$probit, Data$logit) ~ YearC,
    family = binomial("logit"), data = Data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.621080   0.050908   12.20   <2e-16 ***
YearC       -0.025317   0.001224  -20.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 626.91 on 54  degrees of freedom
Residual deviance: 194.35 on 53  degrees of freedom
AIC: 509.25


Number of Fisher Scoring iterations: 3
```

1. probit

```
Call:

glm(formula = = cbind(Data$probit, Data$logit) ~ YearC,
    family = binomial("probit"), data=Data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3856193  0.0317087   12.16   <2e-16 ***
YearC       -0.0157446  0.0007603  -20.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 626.91 on 54  degrees of freedom
Residual deviance: 194.79 on 53  degrees of freedom
AIC: 509.69

Number of Fisher Scoring iterations: 3
```

4 Year Confidence Interval

Calculate an approximate confidence interval for the change in the proportion of papers using logits over time, using the model with the logit link function.

(give your answer to 3 decimal places)

Lower: [ ] .

Upper: [ ] .

Maximum marks: 4

5 What has been going on?

What can you conclude about how logits and probits have been used, and how this has changed over time?

(you can also use this space to show your working for the previous problem)

Fill in your answer here

Maximum marks: 3

6 Why no p-value?

Why can't we use one of the usual significance tests to get a p-value to test if one model is better than the other?

Fill in your answer here

Maximum marks: 2

7 Which is better?

Although we cannot do a significance test using these outputs, we can still compare the model fits. Which of the models explains the data better, and why do you conclude this?

Fill in your answer here

Maximum marks: 3

8 Overdispersion

Is there evidence that the data are overdispersed in either model? Explain how you come to that conclusion.

Fill in your answer here

Maximum marks: 2

9 Getting the Score (Function)

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

The binomial distribution can be written in this form

$$f(r_i|n_i,p_i) = \frac{n_i!}{r_i!(n_i-r_i)!}p_i^{r_i}(1-p_i)^{r_i}$$

The probit response function is

$$p_i = \Phi(\eta_i) = \frac{1}{2\pi}\int_{-\infty}^{\eta_i} e^{-\frac{1}{2}z^2}dz$$

so the link function is just written as $\eta_i = \Phi^{-1}(p_i)$. We should write this as a member of the exponential family:

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

a. Show that $\theta = \log\frac{\Phi(\eta_i)}{1-\Phi(\eta_i)}$

b. Derive E(Y) and Var(Y) as a function of η

c. Show that the score can be written as

$$s_i(\beta) = \frac{(y_i - \Phi(x_i'\beta))x_i f(x_i'\beta)}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}$$

d. Write down the expression for the asymptotic distribution of the parameters, and explain how it would be calculated by the computer

Maximum marks: 15

10 Getting the MLEs

Explain what you need to derive the maximum likelihood estimates, and outline how you would calculate them.

Fill in your answer here

Maximum marks: 3

The use of logits and probits might be different between different scientific fields, e.g. because a Big Name in the field said that one was better. So we can add Field to the model as a categorical covariate.

After removing the fields that don't use logits or probits enough, there are 26 fields. We add it as a random effect, leading to this model:

glmm1 <- glmer(Resp ~ YearC + (1|Field) + (1|OD), family=binomial("logit"), data=MostData)

where

- **Resp** is the response, with logit as a success and probit as failure.
- **YearC** is the year - 1970,
- **Field** is a factor for the field of study
- **OD** is an overdispersion term, i.e. it is a factor with one level for every year/ Field combination.

This gave the following summary

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Resp ~ YearC + (1 | Field) + (1 | OD)
   Data: MostData
   AIC      BIC   logLik deviance df.resid
3806.6   3827.7  -1899.3   3798.6     1426

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2466 -0.4978  0.0000  0.4824  2.7261

Random effects:
 Groups Name        Variance Std.Dev.
 OD     (Intercept) 0.0194   0.1393
 Field  (Intercept) 0.5646   0.7514
Number of obs: 1430, groups:  OD, 1430; Field, 26

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.030925   0.152663  -0.203    0.839
YearC        0.013580   0.001731   7.846 4.29e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Correlation of Fixed Effects:
      (Intr)
YearC -0.220
```

11 The Equation

Which of these is the equations best represents the linear predictor of the fitted model?
Select one alternative:

- $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j + \epsilon_{ij}$
- $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j x_{ij} + \epsilon_{ij}$
- $\eta_{ij} = \beta x_i + \gamma_j + \epsilon_{ij}$
- $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j$

Maximum marks: 1

12 Intraclass Correlation

We can get some idea of how much variation there is between fields compared to within fields by looking at the intraclass correlation. If S is the between study variance and R is the residual variance (= overdispersion here), how is the intraclass correlation calculated?

S/(S+R)

(if you struggle to use the maths tool, use the answer to a later question or the paper to write it out)

Maximum marks: 1

13 Intraclass Correlation Calculation

We can calculate an intra-class correlation (ICC) as we can in a linear mixed model, although the interpretation is slightly different.

What is the ICC for this model?

0.967 .

Maximum marks: 1

14 Comparison of Effects

Based on this, comment on the importance of the Year, Field, and overdispersion in the data. (note that the standard deviation of Year is 15.9)

Fill in your answer here

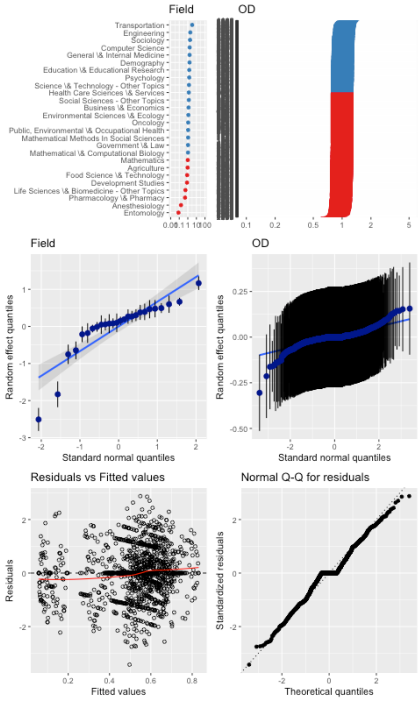
Maximum marks: 4

15 How the Model was fitted

The model fitting produced warnings. We will not look at them, but what are the general problems with fitting a model like this, i.e. why is it more difficult than fitting a GLM or a linear mixed model?

Fill in your answer here

Maximum marks: 2



Here are some plots we can use to assess the fit of the GLMM, from the logit model.

Top row: random effects for Field (left) and overdispersion (right)
Middle row: normal probability plots for Field (left) and overdispersion (right)
Bottom row: residual plot (left), normal probability plot for residuals (right)

16 Model Fit

Comment on how well the model fits the data. Include comments on what assumptions the seem to be reasonable, and what does not.

Fill in your answer here

Maximum marks: 4