Norwegian University of Science and Technology
Department of Mathematical Sciences

English

Contact during exam: Thiago G. Martins 46 93 74 29

EXAM IN COURSE TMA4315 Generalized Linear Models

December 13th 2012
Hours: 09:00–13:00

Permitted aids: *Tabeller og formler i statistikk*, Tapir Forlag
K. Rottmann: *Matematisk formelsamling*
Calculator HP30S / CITIZEN SR-270X
Yellow, stamped A4-sheet with your own handwritten notes.

Examination results are due: January 10th 2013

Problem 1 Precipitation in Trondheim tomorrow?
The NTNU maths student Konrad In-Control, likes to be prepared for the next day, and he wants to know whether there will be precipitation or not tomorrow morning. He constructs statistical models for precipitation occurrence, and consider three explanatory variables;

1. Amount of precipitation (in mm) according to the weather forecast for tomorrow ($Fore$).
2. A binary variable $ForeBin$ that indicates whether the forecast says precipitation ($ForeBin = 1$) or no precipitation ($ForeBin = 0$).
3. A variable OF that indicates how yesterdays observation and forecast were;
   - $OF = 0$: non precipitation observed and non in forecast
   - $OF = 1$: precipitation occurrence observed, and no occurrence in forecast
   - $OF = 2$: no precipitation observed, but occurrence in forecast
   - $OF = 3$: precipitation occurrence observed and occurrence in forecast

Konrad has gathered data for 100 successive days. Data for the ten first days are given in table 1. He used *R* to fit four models (see edited printout below); *model 1* gives `result1`, *model 2* gives `result2` and *model 3* gives `result3`.

```
summary(result1)

Call:
glm(formula = Occur ~ -1 + Fore + ForeBin + as.factor(OF),
family = binomial(link = "logit"), data = OccurData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1714  -0.6614  -0.5975   0.7493   2.2624

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
Fore             0.4609     0.2294   2.009 0.044494 *
ForeBin          0.8883     0.6559   1.354 0.175641
as.factor(OF)0  -1.6327     0.4349  -3.755 0.000174 ***
as.factor(OF)1  -1.1941     0.9690  -1.232 0.217818
as.factor(OF)2  -2.5144     0.6764  -3.717 0.000202 ***
as.factor(OF)3  -1.7609     0.5983  -2.943 0.003249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.63  on 100  degrees of freedom
Residual deviance: 101.09  on  94  degrees of freedom
AIC: 113.09

> summary(result2)

Call:
glm(formula = Occur ~ Fore + ForeBin + OF, family = binomial(link = "logit"),
    data = OccurData)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6438     0.4280  -3.840 0.000123 ***
```

```
Fore          0.5204     0.2224   2.340 0.019286 *
ForeBin       0.7344     0.6396   1.148 0.250832
OF           -0.1355     0.2026  -0.669 0.503589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125.37  on 99  degrees of freedom
Residual deviance: 103.19  on 96  degrees of freedom
AIC: 111.19

> summary(result3)

Call:
glm(formula = Occur ~ Fore, family = binomial(link = "logit"),
    data = OccurData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1246  -0.6984  -0.6146   0.7471   1.8759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5707     0.3152  -4.983 6.27e-07 ***
Fore          0.6683     0.1856   3.601 0.000317 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125.37  on 99  degrees of freedom
Residual deviance: 104.81  on 98  degrees of freedom
AIC: 108.81
```

a) Set up the generalized linear models (GLM) used for each of the models mathematically. Specify and discuss assumptions.
Further, specify the design matrix X for the first 6 observations for each model.
When relevant specify which strategy that is used to ensure identifiability.
Briefly describe the differences between *model 1* and *model 2*.

Occure	Fore	ForeBin	OF
1	3.0	1	3
0	0.5	1	3
0	0.0	0	2
0	0.0	0	0
0	0.0	0	0
0	0.0	0	0
0	0.7	1	0
0	0.4	0	2
0	0.0	0	0
1	0.4	0	0

Table 1: Data from ten days in Konrads's dataset on precipitation occurrence, no occurrence(Occure= 0) or occurrence (Occure = 1). Also available; amount precipitation in forecast in mm, binary forecast and the OF variable.

b) Based on the results from `R` answer the following questions:
According to *model 1*: What is the probability for precipitation if the forecast is 5mm and $OF = 0$?
According to *model 2*: What is the probability for precipitation if the forecast is 5mm and $OF = 3$?
According to *model 3*: What is the odds ratio between a day with forecast 0mm and a day with forecast 5mm?

c) Konrad now wants to compare models: Set up a hypothesis for testing model 1 against model 3 using the likelihood ratio test (i.e. based on deviance), and do the test. Which of the models, model 1, model 2 or model 3, would you prefer. Why?

d) Konrad is also interested in the precipitation occurrence at Trondheim Airport Værnes and the local skiing resort Vassfjellet. He has observations and forecasts also for these locations $Location \in \{\texttt{Trondheim}, \texttt{Vaernes}, \texttt{Vassfjellet}\}$ for the same 100 days.

He fits three models using R running:

```
res4 = glm(Occur~Fore, family=binomial(link="logit"),data=OccurDataTVV)
res5 = glm(Occur~Fore + Location, family=binomial(link="logit"),data=OccurDataTVV)
res6 = glm(Occur~Fore*Location, family=binomial(link="logit"),data=OccurDataTVV)
```

Explain briefly the three models, and make sketches that explains the models.
Consider the used model used to obtain `res4`, and discuss if the assumptions for GLMs are met now. Suggest an alternative model.

Problem 2 Precipitation in Trondheim as snow, sleet or rain

Given that it is precipitation tomorrow, Konrad is also interested in the kind of precipitation tomorrow. We denote this quantity C, and classify precipitation into three classes; snow ($C = 1$), sleet ($C = 2$) and rain ($C = 3$). Further, we want to explore whether the temperature forecast is a good explanatory variable.

a) Let C_i denote the class of precipitation on day i, and let t_i denote the temperature forecast valid for day i.
Suggest an appropriate model for C. Specify the model both mathematically and with words/figure(s). Especially explain how the parameters should be interpreted.

Problem 3 Precipitation in Trondheim, amount

We now want to model the amount of daily precipitation given that it is precipitation, and denote this quantity Y. It is common to model Y as a gamma distributed random variable, $Y \sim Ga(\alpha, \beta)$, with density;

$$f_Y(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta).$$

In this problem we consider N observations, each gamma distributed with $Y_i \sim Ga(\alpha_i, \mu_i/\alpha_i)$. Here α_is are considered known, i.e. they are nuisance parameters.

a) Show that the gamma probability function is member of the exponential family when μ_i is the parameter of interest.
Use this to find expressions for the expected value and the variance of Y_i, in terms of (α_i, μ_i), and interpret α_i.

b) Explain what a saturated model is.
Set up the log-likelihood function expressed by μ_i, and use it to find the maximum likelihood estimators for μ_i-s of the saturated model.
Find the deviance (based on all N observations).

c) We now want to construct a model for amount of precipitation (given that there are occurrence) with precipitation forecast as explanatory variable.

Let Y_i be amount of precipitation for day i, and let x_i be the precipitation forecast valid for day i. Set up a GLM for this, and argue for your choice of link function and linear component.