Norwegian University of Science and Technology
Department of Mathematical Sciences

Contact during exam:
Ingelin Steinsland (92 66 30 96)

EXAM IN TMA4315 GENERALIZED LINEAR MODELS

Friday December 10th, 2010
Time: 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: December 28th 2010

Problem 1 Number of buss passengers

A bus driver wants to model how many passengers he gets from the bus stop close to the student home. He can think of three explanatory variables; which route it is (8 am or 9 am), if it is during the semester or not, and the temperature. He has data for 20 days, given in the table below. He consider three different models, all analyzed in R (see edited printout below); *model 1* gives `result1`, *model 2* gives `result2` and *model 3* gives `result3`.

a) Set up the generalized linear model (GLM) used for *model 1* mathematically, specify assumptions, and specify the design matrix X for the first 6 observations. Also specify which strategy that is used to ensure identifiability, and discuss briefly alternative(s). Explain, mathematically and with words, what model the `R` notation `temp*semester` gives (as in *model 2*).

	Passengers	route	semester	temp
1	3	8am	semester	8.8
2	1	9am	nonSemester	11.5
3	1	8am	nonSemester	12.0
4	3	8am	semester	14.8
5	0	8am	nonSemester	-1.2
6	0	8am	nonSemester	7.8
7	0	8am	nonSemester	6.9
8	1	9am	nonSemester	7.5
9	6	8am	semester	7.7
10	2	8am	semester	5.5
11	1	8am	nonSemester	13.7
12	1	8am	nonSemester	13.1
13	0	9am	nonSemester	14.2
14	2	9am	nonSemester	0.2
15	4	8am	nonSemester	-4.7
16	0	9am	nonSemester	26.3
17	3	9am	semester	3.1
18	2	8am	semester	-4.0
19	1	9am	nonSemester	18.4
20	2	8am	nonSemester	-5.0

b) Consider *model 1*. Based on the results from `R`:
What is the expected number of passengers for the 9 am route, during the semester when it is 5.4 degrees C?
What is the expected number of passengers for the 8 am route, during non-semester when it is -15.2 degrees C?

c) We now want to compare models: Set up a hypothesis for testing *model 2* against *model 1* 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) Let $Y_1, \dots Y_N$ be independent responses with $Y_i \sim Po(\lambda_i)$. For the model of interest, with $p < N$ parameters, let $\hat{y}_i$ be the fitted values based on the maximum likelihood estimates. Find an expression, based on y_i and $\hat{y}_i$, for the deviance in this case.

```
> result1 = glm(Passengers~temp+semester, family=poisson(link="log"))
> summary(result1)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.25406    0.30667   0.828  0.40741
temp             -0.03451    0.02462  -1.401  0.16107
semestersemester  1.08499    0.35365   3.068  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 17.677  on 17  degrees of freedom
AIC: 62.03

> result2 = glm(Passengers~temp*semester, family=poisson(link="log"))
> summary(result2)
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.44315    0.29124   1.522   0.1281
temp                  -0.07445    0.03384  -2.200   0.0278 *
semestersemester       0.54611    0.46383   1.177   0.2390
temp:semestersemester  0.10002    0.05316   1.881   0.0599 .

    Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 13.981  on 16  degrees of freedom
AIC: 60.334

> result3 = glm(Passengers~temp+semester+route, family=poisson(link="log"))
> summary(result3)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.28227    0.32780   0.861  0.38918
temp             -0.03345    0.02501  -1.338  0.18095
semestersemester  1.06849    0.36035   2.965  0.00303 **
route9am         -0.09713    0.42224  -0.230  0.81806

   Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 17.623  on 16  degrees of freedom
AIC: 63.976
```

Problem 2 Negative binomial distribution

The probability density function for a negative binomial random variable is

$$f_y(y;\theta,r) = \frac{\Gamma(y+r)}{y!\Gamma(r)}(1-\theta)^r\theta^y$$

for $y = 0, 1, 2, \ldots$, $r > 0$ and $\theta \in (0, 1)$, and where $\Gamma()$ denotes the gamma function. (There are also other parameterizations of the negative binomial distributions, but use this for now.)

a) Show that the negative binomial distribution is a member of the exponential family. You can in this question consider r as a known constant.

b) Use the general formulas for a exponential family to show that $E(Y) = \mu = r\frac{\theta}{1-\theta}$ and $Var(Y) = \mu\frac{1}{1-\theta}$

c) Set up a GLM for the dataset in problem 1 with a negative binomial response function and a linear component similar to that in *model 1.*
Argue for your choice of link-function.
What role does r have?
In which situations could it be beneficial to use a negative binomial response function instead of a Poisson response function? Why?