Norwegian University of Science and Technology
Department of Mathematical Sciences

English

Contact during exam:
Ingelin Steinsland 926 630 96

EXAM IN COURSE TMA4315 Generalized Linear Models

December 6th 2011
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: December 28th 2011

Problem 1 Christmas gift preferences

Santa wants to model gift preferences; whether children want soft or hard gifts. He has two explanatory variables; age (given in years) and sex (male or female), and data from 100 children, 10 of them is given in Table 1. He considers 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).

b) Explain, mathematically and with words, the three different models. In paticular explain what model the `R` notation `age*sex` gives (as in *model 3*). Further, make a sketch that graphically illustrates the differences between the three models.

	pref	**sex**	**age**	**agegr**
1	1	female	7.5	3
2	0	male	10.5	3
3	0	female	2.0	1
4	1	female	8.0	3
5	0	male	2.0	1
6	1	female	4.0	2
7	1	female	6.0	2
8	1	male	10.0	3
9	1	female	8.0	3
10	0	female	3.0	1

Table 1: Data from ten children in Santa's dataset on preference of soft ($pref = 0$) or hard ($pref = 1$) Christmas gifts. Also available; *age* in years, *sex* and age group *agrgr* as defined in d)

```
> summary(result1)

Call:
glm(formula = pref ~ sex + age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5356  -1.1459   0.8824   1.0555   1.4217

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.69613    0.18356  -3.792 0.000149 ***
sexmale      0.78254    0.13010   6.015  1.8e-09 ***
age          0.06906    0.02529   2.731 0.006311 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1338.0  on 997  degrees of freedom
AIC: 1344


> summary(result2)

Call:
```

```
glm(formula = pref ~ age, family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.353  -1.196   1.026   1.129   1.267

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.35167    0.17067  -2.061  0.03935 *
age          0.07195    0.02483   2.898  0.00376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1374.9  on 998  degrees of freedom
AIC: 1378.9


> summary(result3)

Call:
glm(formula = pref ~ sex * age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5287  -1.1495   0.8866   1.0504   1.4280

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.716715   0.237113  -3.023  0.00251 **
sexmale      0.826983   0.348833   2.371  0.01775 *
age          0.072290   0.034541   2.093  0.03636 *
sexmale:age -0.006965   0.050707  -0.137  0.89076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1338.0  on 996  degrees of freedom
AIC: 1346
```

c) Consider *model 1*. Based on the results from `R`:
What is the probability that a 5 years old girl prefer soft gifts? What is the probability that a 15 years old girl prefer soft gifts? What is the odds ratio between a 15 years old girl and a 15 year old boy for preferring soft gifts?

Based on the evaluation of the results from models 1-3 and his experience, Santa also want to try a model where age is consider a factor with three levels; *agegr1* (0-3 years), `agegr2` (3.5-7 years), `agegr3` (>7). He fits one model, *model 4* which gives `result4`.

d) Based on the results for model 1-4, how can you compare and evaluate the models. Demonstrate your method (for at least one model / one pair of models).
Describe how you can can proceed evaluating the models graphically.
Also suggest a model that you would try to fit next. Explain why.

```
 > summary(result4)

Call:
glm(formula = pref ~ sex + as.factor(agegr), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6569  -1.0654   0.7645   1.0962   1.9477

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.7341     0.2254  -7.694 1.43e-14 ***
sexmale             0.8865     0.1367   6.486 8.81e-11 ***
as.factor(agegr)2   1.9281     0.2338   8.247  < 2e-16 ***
as.factor(agegr)3   1.3020     0.2346   5.550 2.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 1383.4  on 999  degrees of freedom
Residual deviance: 1261.8  on 996  degrees of freedom
AIC: 1269.8
```

e) Santa's department of customer relationships have registered that more and more children want electronic gifts. How can your favorite model among *model 1-4* be extended to account for this. Suggest a model, and discuss your choices.

Problem 2 Number of Christmas gifts

Santa has a model he samples from to get the number of gifts he gives to each child. Let us assume it is Poisson.

a) Show that the Poisson distribution is member of the exponential family, and use this to find the expected value and variance for Poisson distributed random variables.

We have data on number of gifts (Y_i) and age (x_i) for $n = 20$ children. Assume that $Y_1, \ldots, Y_n$ are Poisson with expected value $\mu_i = \exp(\alpha + \beta x_i)$ where α and β are regression parameters.

b) Set up the log-likelihood function for the data described above.
Show that the score function can be written as

$$U(\alpha, \beta) = \begin{pmatrix} U_1(\alpha, \beta) \\ U_2(\alpha, \beta) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (Y_i - \mu_i)$$

Also find an expression for the expected information matrix.

c) What is a saturated model?
Show that the maximum likelihood estimates for the saturated model is $\tilde{\mu}_i = Y_i$ for $i = 1, 2, \ldots, 20$

d) Based on the results above, write an expression for the deviance for this model.

Below is an edited printout from `R` analyzing these data. What is the missing number / degrees of freedom?

Further explain how deviance can be used for this kind of models (generalized linear models with Poisson likelihood).

```
> summary(resut5)

Call:
glm(formula = gifts ~ age, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8766  -0.9863  -0.4624   0.6776   1.7126

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8952     0.4291   4.417    1e-05 ***
age          -0.3323     0.0935  -3.555 0.000379 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 35.941  on 19  degrees of freedom
Residual deviance: 19.805  on ??  degrees of freedom
AIC: 53.599
```

Problem 3 Valid GLMs

Define the class of generalized linear models (GLMs), and explicitly list all requirements for each part of the model.

Below are three likelihoods, three link functions and three linear components listed. Explain which combinations that give valid GLMs, and also comment on these models. (You do not have to mathematically prove which are valid).

Likelihoods:
1. Gaussian; $Y \sim N(\mu, \sigma^2)$
2. Multinomial; $Y \sim M([p_1, p_2, p_3, p_4], N)$
3. Poisson; $Y \sim Po(\theta)$

Link functions:
1. $\eta = \cos(\mu)$
2. $\eta = \mu$
3. $\eta = \log(\mu)$

Linear component:
1. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
2. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$
3. $\eta = \beta_0 + \beta_1 x_1 + \beta_1^2 x_2$