UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to generalized linear models.
Day of examination:	Wednesday 20 December 2017.
Examination hours:	09.00 – 13.00.
This problem set consists of 4 pages.	
Appendices:	Formulas in STK3100/4100.
Permitted aids:	Approved calculator and collection of formulas for STK1100/STK1110 and STK2120.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

Assume that the random variable Y is Poisson distributed with probability mass function (pmf)

$$P(Y = y \,|\, \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda), \quad y = 0, 1, 2, \ldots. \tag{1}$$

a) Show that the distribution of Y is in the exponential dispersion family. That is, show that (1) can be written on the form

$$\exp\{[\theta y - b(\theta)]/a(\phi) + c(y, \phi)\}, \tag{2}$$

and determine θ, $b(\theta)$, $a(\phi)$ and $c(y, \phi)$.

We then assume that $Y_1, Y_2, \ldots, Y_n$ are independent with pmf of the form (1), and let $\mu_i = E(Y_i)$; $i = 1, \ldots, n$.

b) Explain what we mean by a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_n$ with link function g, and determine the canonical link function.

c) Derive an expression for the log-likelihood function $L(\boldsymbol{\mu}; \boldsymbol{y})$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the observed value of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$.

d) Explain what we mean by a saturated model and determine the maximum of $L(\boldsymbol{\mu}; \boldsymbol{y})$ for the saturated model.

e) Explain what we mean by the deviance $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ of a Poisson GLM, find an expression for the deviance, and discuss how it may be used.

(Continued on page 2.)

Problem 2

We assume that the random variable Λ is gamma distributed with pdf

$$f(\lambda; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} \lambda^{k-1} e^{-k\lambda/\mu}\,; \;\; \lambda > 0,$$

and further that given $\Lambda = \lambda$, the random variable Y is Poisson distributed with parameter λ. Thus the conditional pmf of Y given $\Lambda = \lambda$ takes the form (1).

a) Show that the marginal pmf of Y is given by

$$p(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k \;; \;\; y = 0, 1, 2, \ldots$$

This is the negative binomial distribution.

We then assume that the parameter k is fixed, and consider the random variable $Y^* = Y/k$. Note that

$$P(Y^* = y^*) = P(Y = ky^*) \quad \text{for} \quad y^* = 0, \frac{1}{k}, \frac{2}{k}, \ldots$$

so Y^* has pmf

$$p^*(y^*; \mu, k) = \frac{\Gamma(ky^*+k)}{\Gamma(k)\Gamma(ky^*+1)} \left(\frac{\mu}{\mu+k}\right)^{ky^*} \left(\frac{k}{\mu+k}\right)^k \;; \;\; y^* = 0, \frac{1}{k}, \frac{2}{k}, \ldots \tag{3}$$

b) Show that (3) is a distribution in the exponential dispersion family (2), with $\theta = \log[\mu/(\mu+k)]$, $b(\theta) = -\log(1-e^\theta)$, and $a(\phi) = 1/k$.

c) Use the expressions for $b(\theta)$ and $a(\phi)$ to determine $E(Y^*)$ and $\text{var}(Y^*)$. Show that $E(Y) = \mu$ and find $\text{var}(Y)$.

Problem 3

In this problem we will consider data from a sociological study of a sample of aboriginal and non-aboriginal children performed in Australia in the 1970s. The children were selected from four age groups (final grade in primary schools and first, second and third form in secondary schools), and the children in each age group were classified as slow or average learners. For the analyses presented in this problem, we use the number of days a child was absent from school during one school year (`Days`) as response. The covariates are all categorical, and they are given as follows:

- `Eth`: Ethnic background (A: aboriginal; N: non-aboriginal)
- `Sex`: Sex (F: girl; M: boy)
- `Age`: age group (F0: primary; F1, F2, F3: first, second and third grade in secondary school)
- `Lrn`: learner status (AL: average learner; SL: slow learner)

(Continued on page 3.)

a) Below is given the result of an analysis of the data. Describe the model that we have used in this analysis, and discuss the assumptions for this model.

```
Call:    glm(formula = Days~Eth+Sex+Age+Lrn, family = "poisson")

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.71538    0.06468  41.980  < 2e-16
EthN        -0.53360    0.04188 -12.740  < 2e-16
SexM         0.16160    0.04253   3.799 0.000145
AgeF1       -0.33390    0.07009  -4.764 1.90e-06
AgeF2        0.25783    0.06242   4.131 3.62e-05
AgeF3        0.42769    0.06769   6.319 2.64e-10
LrnSL        0.34894    0.05204   6.705 2.02e-11
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1696.7  on 139  degrees of freedom
AIC: 2299.2
```

b) The fit of another model is given below. Describe the model that we have used here. Would you prefer this analysis to the one given in question a? Give the reasons for your answer.

```
Call:  glm.nb(formula = Days~Eth+Sex+Age+Lrn, init.theta = 1.275, link = log)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.89458    0.22842  12.672  < 2e-16
EthN        -0.56937    0.15333  -3.713 0.000205
SexM         0.08232    0.15992   0.515 0.606710
AgeF1       -0.44843    0.23975  -1.870 0.061425 .
AgeF2        0.08808    0.23619   0.373 0.709211
AgeF3        0.35690    0.24832   1.437 0.150651
LrnSL        0.29211    0.18647   1.566 0.117236
(Dispersion parameter for Negative Binomial(1.2749) family taken to be 1)
    Null deviance: 195.29  on 145  degrees of freedom
Residual deviance: 167.95  on 139  degrees of freedom
AIC: 1109.2

              Theta:  1.275
          Std. Err.:  0.161
 2 x log-likelihood:  -1093.151
```

c) Finally we consider a model with interaction between ethnic group and age. The results for this model are given on the next page. Explain why you will prefer this model to the one in question b.

(Continued on page 4.)

```
Call:
glm.nb(formula=Days~Eth+Sex+Age+Lrn+Eth:Age, init.theta=1.380, link=log)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.53409    0.27387   9.253  < 2e-16
EthN         0.05698    0.34289   0.166  0.86802
SexM         0.11275    0.15492   0.728  0.46673
AgeF1        0.08732    0.32622   0.268  0.78895
AgeF2        0.70638    0.31878   2.216  0.02670
AgeF3        0.40050    0.33756   1.186  0.23544
LrnSL        0.22754    0.18046   1.261  0.20735
EthN:AgeF1  -0.89843    0.43635  -2.059  0.03950
EthN:AgeF2  -1.18060    0.44357  -2.662  0.00778
EthN:AgeF3  -0.10128    0.46025  -0.220  0.82584
(Dispersion parameter for Negative Binomial(1.3799) family taken to be 1)
    Null deviance: 208.33  on 145  degrees of freedom
Residual deviance: 168.08  on 136  degrees of freedom
AIC: 1104.7

              Theta:  1.380
          Std. Err.:  0.178
 2 x log-likelihood:  -1082.688
```

d) Use the model in question c to describe which effects ethnic group and age have for the expected number of days a child is absent from school.

Problem 4

Assume that U_i is $N(0, \sigma^2)$-distributed and that given $U_i = u_i$, the binary random variables $Y_{i1}, \ldots, Y_{id}$ are independent with

$$P(Y_{ij} = 1 \,|\, U_i = u_i) = 1 - P(Y_{ij} = 0 \,|\, U_i = u_i) = \Phi(\beta_0 + \beta_1 x_{ij} + u_i). \tag{4}$$

Here Φ is the cumulative standard normal distribution, and the x_{ij}'s are known numbers.

a) What is model (4) called? Describe one or more situations where such a model may be useful.

A marginal model for the Y_{ij}'s is given by

$$P(Y_{ij} = 1) = 1 - P(Y_{ij} = 0) = \Phi(\gamma_0 + \gamma_1 x_{ij}). \tag{5}$$

b) Show how the parameters γ_0 and γ_1 in (5) may be expressed in terms of β_0, β_1 and σ^2.

c) Discuss what you may learn from the result in question c when it comes to fitting marginal and random effects models for clustered binary data.

END