UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to generalized linear models. SOLUTIONS TO PROBLEMS
Day of examination:	Wednesday 20 December 2017.
Examination hours:	09.00 – 13.00.

This problem set consists of 8 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

The random variable Y is Poisson distributed with pmf

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

a) We may rewrite (1) as

$$P(Y = y \mid \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda) = \exp\{y \log(\lambda) - \lambda - \log(y!)\},$$

which is on the form

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

with $\theta = \log(\lambda)$, $b(\theta) = \lambda = e^\theta$, $a(\phi) = 1$ and $c(y, \phi) = -\log(y!)$.

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

b) A GLM for $Y_1, Y_2, \dots, Y_n$ with link function g, is specified by assuming that

- $Y_1, Y_2, \dots, Y_n$ are independent and all have pmf on the form (2), which in our case is the same as (1).
- Corresponding to each Y_i we have covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, often with $x_{i1} = 1$ for all i, and a linear predictor $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} = \sum_{j=1}^p \beta_j x_{ij}$.
- The mean $\mu_i = E(Y_i)$ is linked with the linear predictor by the relation $g(\mu_i) = \eta_i$. Here the link function g is a strictly increasing, differentiable function.

(Continued on page 2.)

We have a canonical link function when the linear predictor η_i is equal to the natural parameter θ_i, i.e. when $g(\mu_i) = \theta_i$. From question a we have that

$$\log(\mu_i) = \log(\lambda_i) = \theta_i,$$

so $g(\mu_i) = \log(\mu_i)$ is the canonical link function.

c) We have that $Y_1, Y_2, \ldots, Y_n$ are independent with pmf of the form (1) with $\lambda_i = \mu_i$. Therefore the likelihood function is given by

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i).$$

Hence the log-likelihood function becomes

$$L(\boldsymbol{\mu}; \boldsymbol{y}) = \log\{\ell(\boldsymbol{\mu}; \boldsymbol{y})\} = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i - \log(y_i!)\}.$$

d) For a saturated model there are no restrictions on the expected values, so there is a separate parameter μ_i for each observation y_i.

The log-likelihood obtains its maximum value when

$$\frac{\partial}{\partial \mu_i} L(\boldsymbol{\mu}; \boldsymbol{y}) = 0 \quad \text{for all } i = 1, \ldots, n.$$

Now we have

$$\frac{\partial}{\partial \mu_i} L(\boldsymbol{\mu}; \boldsymbol{y}) = \frac{y_i}{\mu_i} - 1\,,$$

so the log-likelihood takes its maximal value when $y_i/\mu_i - 1 = 0$. Thus the ML estimates for the saturated model are $\widetilde{\mu}_i = y_i$, and the maximal value of the log-likelihood becomes

$$L(\boldsymbol{y}; \boldsymbol{y}) = \sum_{i=1}^{n} \{y_i \log(y_i) - y_i - \log(y_i!)\}.$$

e) For a Poisson GLM we have $a(\phi) = 1$; cf. question a. Then the deviance $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ for a model with fitted values $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^{\mathrm{T}}$ is given as

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2 \log \left(\frac{\text{max likelihood for actual model}}{\text{max likelihood for saturated model}} \right).$$

The deviance measures how far the log-likelihood of the model is from the maximum value of of the log-likelihood. For a Poisson GLM the deviance is given by

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2 \log \left(\frac{\prod_{i=1}^{n} (\hat{\mu}_i^{y_i}/y_i!) \exp(-\hat{\mu}_i)}{\prod_{i=1}^{n} (y_i^{y_i}/y_i!) \exp(-y_i)} \right) = 2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i \right\}.$$

(Continued on page 3.)

The deviances may be used for comparing nested models. In order to explain how this may be done, we consider two Poisson GLM models with the same link function g. For model M_1 we have the linear predictors $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_{ij}$; $i = 1, \ldots, n$, while the linear predictors for model M_0 are obtained by setting $p-q$ of the β_j's equal to zero (or by imposing $p-q$ linear restrictions on the β_j's). Thus model M_0 has q parameters. The fitted values under model M_0 and M_1 are denoted, respectively, $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$

We now assume that model M_1 holds and want to test the null hypothesis that also model M_0 holds. The likelihood ratio test for this hypothesis problem rejects the null hypothesis for large values of

$$\begin{aligned} G^2(M_0 \,|\, M_1) &= -2\log\left(\frac{\text{max likelihood for model } M_0}{\text{max likelihood for model } M_1}\right) \\ &= -2\log\left(\frac{\text{max likelihood for actual model}}{\text{max likelihood for saturated model}}\right) \\ &\qquad +2\log\left(\frac{\text{max likelihood for actual model}}{\text{max likelihood for saturated model}}\right) \\ &= D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) \end{aligned}$$

Thus the difference between the deviances of the two nested models M_0 and M_1 can be used for testing the null hypothesis that model M_0 holds. When model M_0 holds, we have that the difference between the deviances is approximately chi-squared distributed with $p-q$ degrees of freedom.

Problem 2

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

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

and further that given $\Lambda = \lambda$, the conditional pmf of the random variable Y, given $\Lambda = \lambda$, takes the form (1).

a) For $y = 0, 1, \ldots$, the marginal pmf of Y is given by

$$\begin{aligned} p(y; \mu, k) &= P(Y = y \,|\, \mu, k) \\ &= \int_0^\infty P(Y = y \,|\, \lambda)\, f(\lambda; k, \mu)\, d\lambda \\ &= \int_0^\infty \frac{\lambda^y}{y!} \exp(-\lambda) \frac{(k/\mu)^k}{\Gamma(k)} \lambda^{k-1} e^{-k\lambda/\mu}\, d\lambda \\ &= \frac{(k/\mu)^k}{\Gamma(k) y!} \int_0^\infty \lambda^{y+k-1} e^{-(\mu+k)\lambda/\mu}\, d\lambda \\ &= \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)} \int_0^\infty \left(\frac{\mu}{\mu+k} u\right)^{y+k-1} e^{-u} \frac{\mu}{\mu+k}\, du \qquad [\text{substitute } u = (\mu+k)\lambda/\mu] \end{aligned}$$

(Continued on page 4.)

$$
\begin{aligned}
&= \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)}\left(\frac{\mu}{\mu+k}\right)^{y+k}\int_0^\infty u^{y+k-1}e^{-u}\,du \\
&= \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)}\left(\frac{\mu}{\mu+k}\right)^{y+k}\Gamma(y+k) \\
&= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{\mu}{\mu+k}\right)^{y}\left(\frac{k}{\mu+k}\right)^{k}.
\end{aligned}
$$

We now assume that the parameter k is fixed, and consider the random variable $Y^* = Y/k$. We have that $P(Y^* = y^*) = P(Y = ky^*)$, so Y^* has pmf

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

b) If we introduce

$$
c(y^*,k) = \log\left(\frac{\Gamma(ky^*+k)}{\Gamma(k)\Gamma(ky^*+1)}\right),
$$

we may rewrite the pmf (3) as follows

$$
\begin{aligned}
p^*(y^*;\mu,k) &= \exp\left\{(ky^*)\log\left(\frac{\mu}{\mu+k}\right) + k\log\left(\frac{k}{\mu+k}\right) + c(y^*,k)\right\} \\
&= \exp\left\{\left[y^*\log\left(\frac{\mu}{\mu+k}\right) + \log\left(\frac{k}{\mu+k}\right)\right]\Big/\frac{1}{k} + c(y^*,k)\right\} \\
&= \exp\left\{\left[y^*\log\left(\frac{\mu}{\mu+k}\right) + \log\left(1-\frac{\mu}{\mu+k}\right)\right]\Big/\frac{1}{k} + c(y^*,k)\right\}.
\end{aligned}
$$

This is of the form (2), with

$$
\begin{aligned}
&\theta = \log\left(\frac{\mu}{\mu+k}\right), \\
&b(\theta) = -\log\left(1-\frac{\mu}{\mu+k}\right) = -\log(1-e^\theta), \\
&a(\phi) = \frac{1}{k}.
\end{aligned}
$$

c) From general results for the exponential dispersion family, we have that $E(Y^*) = b'(\theta)$ and $\text{var}(Y^*) = a(\phi)b''(\theta)$. Hence we have that

$$
E(Y^*) = \frac{d}{d\theta}\left[-\log(1-e^\theta)\right] = \frac{e^\theta}{1-e^\theta} = \frac{\mu/(\mu+k)}{1-\mu/(\mu+k)} = \frac{\mu}{k}.
$$

(Continued on page 5.)

and

$$\begin{aligned}\mathrm{var}(Y^*) &= \frac{1}{k}\,\frac{d^2}{d\theta^2}\left[-\log(1-e^\theta)\right] = \frac{1}{k}\,\frac{e^\theta}{(1-e^\theta)^2}\\ &= \frac{1}{k}\,\frac{\mu/(\mu+k)}{[1-\mu/(\mu+k)]^2} = \frac{1}{k}\,\frac{\mu/(\mu+k)}{[k/(\mu+k)]^2}\\ &= \frac{1}{k^3}\,\mu(\mu+k).\end{aligned}$$

Now $Y = kY^*$, so we have

$$E(Y) = kE(Y) = k\,\frac{\mu}{k} = \mu,$$

and

$$\mathrm{var}(Y) = k^2\,\mathrm{var}(Y^*) = \frac{1}{k}\,\mu(\mu+k) = \mu + \frac{\mu^2}{k}.$$

Problem 3

a) The analysis reported in question a is based on a Poisson GLM with log link. To describe the model we let Y_i denote the number of days absent from school for child number i, and we let $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{i6})$ denote its covatiates (including $x_{i0} = 1$ for the intercept):

$x_{i1} = 1$ if child i is non-aboriginal; $x_{i1} = 0$ if child i is aboriginal,

$x_{i2} = 1$ if child i is a boy; $x_{i2} = 0$ if child i is a girl,

$x_{i3} = 1$ if child i is in first form in secondary school; $x_{i3} = 0$ otherwise,

$x_{i4} = 1$ if child i is in second form in secondary school; $x_{i4} = 0$ otherwise,

$x_{i5} = 1$ if child i is in third form in secondary school; $x_{i5} = 0$ otherwise,

$x_{i6} = 1$ if child i is a slow learner; $x_{i6} = 0$ if child i is an average learner.

The model assumes that the Y_i's are independent and Poisson distributed with means $\mu_i = E(Y_i)$ given as

$$\mu_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta}) = \exp\left(\sum_{j=0}^{6} \beta_j x_{ij}\right).$$

An implication of the Poisson assumption is that $\mathrm{var}(Y_i)$ is also given by the expression above, and this may be a restrictive assumption. The large residual deviance seen in the output for the Poisson GLM indicate that there is overdispersion in these data, i.e. a dispersion that is larger than predicted by the Poisson model.

(Continued on page 6.)

b) The analysis reported in question b is based on a negative binomial GLM with log link. For this model we still assume that $E(Y_i) = \mu_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta})$, i.e. as in question a. But here $\text{var}(Y_i) = \mu_i + \mu_i^2/k$ is allowed to be larger than the mean (so the model allows for overdispersion).

The AIC for the Poisson model in question a is 2299.2, while the AIC for the negative binomial model is 1109.2. This is a very large reduction in the AIC, so the negative binomial model fits the data much better than the Poisson model.

c) We here consider a negative binomial GLM with interaction between ethnic group and age. So in addition to the covariates in question a, we here also have the covariates

$x_{i7} = 1$ if child i is non-aboriginal and is in first form in secondary school; $x_{i7} = 0$ otherwise,

$x_{i8} = 1$ if child i is non-aboriginal and is in second form in secondary school; $x_{i8} = 0$ otherwise,

$x_{i9} = 1$ if child i is non-aboriginal and is in third form in secondary school; $x_{i9} = 0$ otherwise,

and the expression for $\mu_i = E(Y_i)$ now takes the form

$$\mu_i = \exp\left(\sum_{j=0}^{9} \beta_j x_{ij}\right).$$

The AIC for the model in question b is 1109.2, while it is 1104.7 for the model in question c. So according to AIC, the model in question c should be preferred.

Alternatively, we may use the likelihood ratio test and check if the interaction is significant. From the output we have that

$$\begin{aligned}&-2(\text{likelihood rato})\\ &\quad= -2\log(\text{max likelihood model in b}) + 2\log(\text{max likelihood model in c})\\ &\quad= 1093.151 - 1082.688 = 10.463\end{aligned}$$

This should be compared to a chi-squared distribution with three degrees of freedom, which give a P-value of 1.5%. (As tables were not provided, the students could not compute the P-value at the exam.) Thus the interaction is significant, so we should prefer the model in question c.

d) Sex and learner status do not enter in any interactions, so they will have a proportional effect on the estimates for the expected number days a child is absent from school. So when studying the effects ethnic group and age, we may consider the reference levels of sex (which is girl) and learner status (which is average learner).

With sex and learning status at their reference levels, the expected number of days absent for an aboriginal child is estimated to be

Final grade in primary school: $\exp(2.534) = 12.6$

(Continued on page 7.)

First form in secondary school: $\exp(2.534 + 0.087) = 13.7$

Second form in secondary school: $\exp(2.534 + 0.706) = 25.5$

Third form in secondary school: $\exp(2.534 + 0.401) = 18.8$

while for a non-aboriginal child we obtain the estimates

Final grade in primary school: $\exp(2.534 + 0.057) = 13.3$

First form in secondary school: $\exp(2.534 + 0.057 + 0.087 - 0.898) = 5.9$

Second form in secondary school: $\exp(2.534 + 0.057 + 0.706 - 1.181) = 8.3$

Third form in secondary school: $\exp(2.534 + 0.057 + 0.401 - 0.101) = 18.0$

We see that the expected number of days absent are about the same for the ethnic groups for the children in final grade in primary school and for third form in secondary school. But for first and second form in secondary school, an an aboriginal child may expect more than two times as many days of absence as a non-aboriginal child.

Problem 4

We 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

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

a) The model (4) is a generalized linear mixed model (GLMM). More specifically, it is a probit-normal model for binary data with random intercept. The model may be used to study clustered binary data, e.g. the occurrence of a disease in litters of test animals (each litter is a cluster) or the responses to a number of related yes/no questions for a number of people (the answers for one person constitute a cluster). The effect of the random intercept u_i is to make the observations for the units in a cluster correlated.

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

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

b) In order to study the relation between the GLMM model and the marginal model, we will derive the marginal probability corresponding to (4). To this end we let Z be a standard normal random variable that is independent of U_i, and note that since $\Phi(z) = P(Z \le z)$, we may write

$$\begin{aligned} P(Y_{ij} = 1 \,|\, U_i = u_i) &= \Phi(\beta_0 + \beta_1 x_{ij} + u_i) \\ &= P(Z \le \beta_0 + \beta_1 x_{ij} + u_i) \\ &= P(Z - u_i \le \beta_0 + \beta_1 x_{ij}). \end{aligned}$$

(Continued on page 8.)

If we let $f_{U_i}(u_i)$ denote the density of U_i, we have that

$$\begin{aligned} P(Y_{ij}=1) &= \int_{-\infty}^{\infty} P(Y_{ij}=1\,|\,U_i=u_i)\,f_{U_i}(u_i)\,du_i \\ &= \int_{-\infty}^{\infty} P(Z-u_i \leq \beta_0+\beta_1 x_{ij})\,f_{U_i}(u_i)\,du_i \\ &= \int_{-\infty}^{\infty} P(Z-U_i \leq \beta_0+\beta_1 x_{ij}\,|\,U_i=u_i)\,f_{U_i}(u_i)\,du_i \\ &= P(Z-U_i \leq \beta_0+\beta_1 x_{ij}). \end{aligned}$$

Now $Z-U_i \sim N(0,1+\sigma^2)$, and therefore

$$\frac{Z-U_i}{\sqrt{1+\sigma^2}} \sim N(0,1).$$

If follows that

$$P(Y_{ij}=1) = P\left(\frac{Z-U_i}{\sqrt{1+\sigma^2}} \leq \frac{\beta_0+\beta_1 x_{ij}}{\sqrt{1+\sigma^2}}\right) = \Phi\left(\frac{\beta_0+\beta_1 x_{ij}}{\sqrt{1+\sigma^2}}\right).$$

If we compare the last equation with (5), we see that the relation between the parameters of the marginal model and the GLMM is given by

$$\gamma_j = \frac{\beta_j}{\sqrt{1+\sigma^2}} \qquad \text{for } j=0,1.$$

c) The interpretation of the regression coefficient γ_1 for the marginal model and the regression coefficient β_1 for the GLMM are not the same. γ_1 is the population effect (on the probit scale, i.e. the scale of the linear predictor) of one unit's increase in the covariate x_1 without consideration of clusters, while β_1 is the effect of one unit's increase of the covariate when considering two units from the same cluster (or with the same value of the random intercept).

From the result in question b, we see that the regression coefficient γ_1 of the marginal model is closer to zero than the regression coefficient β_1 of the GLMM. The ratio of the regression coefficients for the GLMM and the marginal model is $\gamma_1/\beta_1 = \sqrt{1+\sigma^2}$. So the larger the variation of the random intercept in the GLMM, the more the two regression coefficients will differ.