UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100 — Introduction to generalised linear models
Day of examination:	Wednesday December 4th, 2024
Examination hours:	09.00 – 13.00
This problem set consists of 8 pages.	
Appendices:	Formulas in STK3100/4100
Permitted aids:	Approved calculator

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

Problem 1

We assume that the random variable Y follows an inverse Gaussian distribution with parameters $\mu, \sigma > 0$, written as $Y \sim IG(\mu, \sigma)$, and thus has the probability density function (pdf)

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi y^3 \sigma}} \exp\left(-\frac{1}{2\sigma}\frac{(y-\mu)^2}{y\mu^2}\right), \quad y > 0. \tag{1}$$

a

Show that the distribution of Y is in the exponential dispersion family, so that its pdf can be written on the form

$$f(y; \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right),$$

with $\theta = \frac{1}{\mu^2}$, $\phi = \sigma$, $b(\theta) = \frac{2}{\mu} = 2\sqrt{\theta}$ and $a(\phi) = -2\phi$, and determine $c(y, \phi)$.

b

Use the expressions for $b(\theta)$ and $a(\phi)$ to show that $\mathrm{E}(Y) = \mu$ and to find $\mathrm{Var}(Y)$.

Assume in the rest of the problem that the random variables $Y_1, \ldots, Y_n$ are independent with $Y_i \sim IG(\mu_i, \sigma)$, $i = 1, \ldots, n$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the vectors of covariate values, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Now, consider a generalised linear model (GLM) with linear predictor $\eta_i = \mathbf{x}_i\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_{ij}$ and link function $g(\mu_i) = \eta_i$.

(Continued on page 2.)

c

- Explain what a canonical link function is, and find the canonical link function for this particular model.
- Further, show that the likelihood equations for this model can be written as

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, \ldots, p.$$

d

Show that the deviance of this GLM is given by

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}\left(\frac{y_i}{\widehat{\mu}_i^2} - \frac{2}{\widehat{\mu}_i} + \frac{1}{y_i}\right).$$

e

- Find the log-likelihood function of this GLM and show that the maximum likelihood estimator of $\phi = \sigma$ is $\widehat{\phi}_{MLE} = \frac{D(\mathbf{Y};\widehat{\boldsymbol{\mu}})}{n}$.
- This is a biased estimator (you do not need to show that). Why would one typically prefer the estimator $\frac{D(\mathbf{Y};\widehat{\boldsymbol{\mu}})}{n-p}$? (*Hint*: Recall that $D(\mathbf{Y}; \widehat{\boldsymbol{\mu}})/\phi \overset{approx}{\sim} \chi^2_{n-p}$).

Problem 2

The balance of a human being can be affected by several factors, such as height, vision and the type of surface the person is standing on. In this problem, we will consider data from a study of balance, that contains $N = 480$ records of the following explanatory variables:

- `Sex`: Gender of the individual (binary; 1: male, 0: female)
- `Age`: Age of the individual in years (numerical)
- `Height`: Height of the individual in cm (numerical)
- `Weight`: Weigh of the individual in kg (numerical)
- `Surface`: Type of surface the individual is standing on (binary; 1: normal, 0: foam)
- `Vision`: Factor denoting whether the individual has eyes closed, open, or a dome placed over the head (categorical; levels: closed, open, dome; closed is the reference level)

The response variable (`ctsib.low`) is a binary indicator of whether the individual has a low ("1") or high ("0") score on the CTSIB balance scale.

(Continued on page 3.)

a

Output from fitting a GLM with all the main effects described above in R is given below.
- Describe the model that is used, including all necessary assumptions.
- Further, give an interpretation of the estimated β-coefficient for `Height`.

```
Call:
glm(formula = ctsib.low ~ Sex + Age + Height + Weight + Surface
       + Vision, family = binomial(link = logit), data = ctsib)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.244964   3.872836   2.904 0.003690 **
Sex          1.401577   0.516231   2.715 0.006627 **
Age          0.002521   0.024307   0.104 0.917390
Height      -0.096413   0.026837  -3.593 0.000327 ***
Weight       0.043503   0.018002   2.417 0.015665 *
Surface     -3.967515   0.447179  -8.872  < 2e-16 ***
Visiondome   0.363753   0.383218   0.949 0.342516
Visionopen   3.187501   0.416001   7.662 1.83e-14 ***
---

Null deviance: 526.25  on 479  degrees of freedom
Residual deviance: 295.20  on 472  degrees of freedom
AIC: 311.2
```

b

As the covariates `Age` and `Weight` do not seem to have a very significant effect on balance (see the R output below), we try to refit the model, first without `Age`, and then without `Weight`. Some of the entries have been replaced by question marks.
- Give the missing numbers, explaining how you found them.
- Which of the models do you think fits the data the best, and why?

```
> drop1(fit.1,test="LRT")
Single term deletions
Model:ctsib.low ~ Sex + Age + Height + Weight + Surface + Vision
        Df Deviance    AIC     LRT  Pr(>Chi)
<none>       295.20 311.20
Sex      1   302.90 316.90   7.693 0.0055446 **
Age      1   295.21 309.21   0.011 0.9174007
Height   1   308.60 322.60  13.400 0.0002516 ***
Weight   1   301.18 315.18   5.977 0.0144946 *
Surface  1   443.27 457.27 148.062 < 2.2e-16 ***
Vision   2   395.90 407.90 100.701 < 2.2e-16 ***
```

(Continued on page 4.)

```
Analysis of Deviance Table
Model 1: ctsib.low ~ Sex + Height + Surface + Vision
Model 2: ctsib.low ~ Sex + Height + Weight + Surface + Vision
Model 3: ctsib.low ~ Sex + Age + Height + Weight + Surface + Vision
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       474     301.53
2       473      ?      1   6.3131  0.01198 *
3        ?      295.20  1     ?     0.91740
```

c

We also tried fitting the GLM with all the main effects except `Age`, but with the probit link function. The R output from the fitted model is given below, as well as a table with AIC values from the models with logit and probit link functions.
- Based on these results, which link function would you prefer?
- Is there a straightforward interpretation of the estimated β-coefficient for `Height` when using the probit link? How might that affect your choice of link function?

```
glm(formula = ctsib.low ~ Sex + Height + Weight + Surface + Vision,
family = binomial(link = probit), data = ctsib)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.214352   2.162243   2.874 0.004053 **
Sex          0.761065   0.282467   2.694 0.007053 **
Height      -0.053601   0.015019  -3.569 0.000358 ***
Weight       0.024556   0.009985   2.459 0.013925 *
Surface     -2.239426   0.235970  -9.490  < 2e-16 ***
Visiondome   0.219990   0.220935   0.996 0.319386
Visionopen   1.873090   0.230159   8.138 4.01e-16 ***
---
Null deviance: 526.25  on 479  degrees of freedom
Residual deviance: 294.06  on 473  degrees of freedom
AIC: 308.06
```

Link function	**AIC**
logit	309.2134
probit	308.0592

Problem 3

We will now consider binary response variables, that are correlated within clusters. Let Y_{ij} denote the response for subject j in cluster i, $i = 1, \ldots, n$, $j = 1, \ldots, d$, and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})$ be vectors of explanatory variables. A binomial generalised linear mixed model (GLMM) with probit link and random intercept $u_i \sim N(0, \sigma_u^2)$, $i = 1, \ldots, n$, is then given by

$$\Phi^{-1}(\mathrm{E}(Y_{ij}|u_i)) = \Phi^{-1}\left(P(Y_{ij} = 1|u_i)\right) = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i,$$

(Continued on page 5.)

where Φ is the cumulatice distribution function of the standard normal distribution. The marginal model is then given by

$$\mu_{ij} = \mathrm{E}(Y_{ij}) = P(Y_{ij} = 1) = \int \mathrm{E}(Y_{ij}|u_i) f(u_i; \sigma_u^2) du_i.$$

a

- Argue that one may write $P(Y_{ij} = 1|u_i) = P(Z \leq \mathbf{x}_{ij}\boldsymbol{\beta} + u_i|u_i)$, where Z is a standard normal random variable independent of u_i, i.e. $Z \sim N(0, 1)$.
- Further, explain why $Z - u_i \sim N(0, 1 + \sigma_u^2)$.
- Then, show that the marginal model is given by

$$\mu_{ij} = \mathrm{E}(Y_{ij}) = P(Y_{ij} = 1) = \Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1 + \sigma_u^2}}\right).$$

- Finally, comment on the relationship between the fixed effects of the probit link binomial GLMM and the binomial GLM with probit link, i.e. $\mathrm{E}(Y_{ij}) = P(Y_{ij} = 1) = \Phi(\mathbf{x}_{ij}\boldsymbol{\beta})$.

We return our attention to the balance study from **Problem 2**, as it turns out that these are not measurements from 480 different individuals, but rather $d = 12$ measurements on each of the $n = 40$ individuals that participated in the study, corresponding to two repeated measurements of the balance score under each of the 6 different combinations of the covariates `Surface` and `Vision`.

b

- Explain why the binomial GLMM, described above, is a more reasonable model for these data, than the binomial GLMs, that were used in **Problem 2**, in light of the model assumptions made in **2a**.
- Compare the R output on the fixed effects from the fitted GLMM, given below, to the GLM fitted in **2c**. Do these results (approximately) agree with the findings from **3b**? (*Hint*: Note that $\widehat{\sigma}_u^2 = 2.242$).
- And what do the P-values from the Wald-tests tell you about the consequence of ignoring correlation between the responses?

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
Gauss-Hermite Quadrature, nAGQ = 70) [glmerMod]
Family: binomial  ( probit )
Formula: ctsib.low ~ Sex + Height + Weight + Surface + Vision + (1 | Subject)
Data: ctsib
AIC      BIC   logLik deviance df.resid
246.3    279.7   -115.2    230.3      472

Random effects:
Groups  Name        Variance Std.Dev.
Subject (Intercept) 2.242    1.497
Number of obs: 480, groups:  Subject, 40
```

(Continued on page 6.)

```
Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.15270    7.02478   1.730   0.0836 .
Sex          1.71537    0.93437   1.836   0.0664 .
Height      -0.10042    0.04868  -2.063   0.0391 *
Weight       0.03764    0.03227   1.166   0.2434
Surface     -4.01708    0.53719  -7.478 7.55e-14 ***
Visiondome   0.34961    0.29574   1.182   0.2371
Visionopen   3.33993    0.48704   6.858 7.00e-12 ***
```

END

(Continued on page 7.)

APPENDIX: Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\},$$

(Continued on page 8.)

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\text{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \dots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \dots, \eta_n$ are given by $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \dots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right), \quad \text{approximately}$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^n \omega_i \left[y_i\left(\tilde{\theta}_i - \widehat{\theta}_i\right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i)\right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{id})^{\mathrm{T}}$ for $i = 1, \dots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots \epsilon_{id})^{\mathrm{T}} \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2\boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \mid \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100 — Introduction to generalised linear models SKETCH OF SOLUTION
Day of examination:	Wednesday December 4th, 2024
Examination hours:	09.00 – 13.00

This problem set consists of 6 pages.

Appendices:

Permitted aids:

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

Problem 1

a

The pdf (1) can be written as

$$\begin{aligned} f(y;\mu,\sigma) =& \frac{1}{\sqrt{2\pi y^3\sigma}} \exp\left(-\frac{1}{2\sigma}\frac{(y-\mu)^2}{y\mu^2}\right) \\ =& \exp\left(-\frac{1}{2\sigma}\left(\frac{y}{\mu^2}-\frac{2}{\mu}+\frac{1}{y}\right)-\frac{1}{2}\log(2\pi y^3\sigma)\right) \\ =& \exp\left(\frac{\frac{1}{\mu^2}y-\frac{2}{\mu}}{-2\sigma}-\frac{1}{2\sigma y}-\frac{1}{2}\log(2\pi y^3\sigma)\right), \end{aligned}$$

which is on the form

$$f(y;\theta,\phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right),$$

with $\theta = \frac{1}{\mu^2}$ and $\phi = \sigma$, so that $\mu = \frac{1}{\sqrt{\theta}}$, $b(\theta) = \frac{2}{\mu} = 2\sqrt{\theta}$, $a(\phi) = -2\sigma = -2\phi$ and $c(y,\phi) = -\frac{1}{2\sigma y} - \frac{1}{2}\log(2\pi y^3\sigma)$.

b

From **a**, we know that $\mu = \frac{1}{\sqrt{\theta}}$, $b(\theta) = \frac{2}{\mu} = 2\sqrt{\theta}$ and $a(\phi) = -2\sigma$. Further, we know that for an exponetial dispersion family $\mathrm{E}(Y) = b'(\theta)$ and $\mathrm{Var}(Y) = a(\phi)b''(\theta)$. This gives:

$$b'(\theta) = \frac{1}{\sqrt{\theta}} = \mu \quad \text{and} \quad b''(\theta) = -\frac{1}{2\theta^{3/2}} = -\frac{1}{2}\mu^3,$$

so that

$$\mathrm{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = -2\sigma \cdot \left(-\frac{1}{2}\mu^3\right) = \sigma\mu^3.$$

(Continued on page 2.)

c

The canonical link function is the one that equates the linear predictor η_i to the natural parameter θ_i. Hence the canonical link is given by:

$$\eta_i = g(\mu_i) = \theta_i = \frac{1}{\mu_i^2}.$$

Further, we know that the likelihood equations of a GLM are given by:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \ldots, p.$$

From **b**, we know that $\text{Var}(Y_i) = \sigma\mu_i^3$. Moreover, we have $\eta_i = \frac{1}{\mu_i^2}$, so that $\mu_i = \frac{1}{\sqrt{\eta_i}}$ and

$$\frac{\partial \mu_i}{\partial \eta_i} = -\frac{1}{2\eta_i^{3/2}} = -\frac{1}{2}\mu^3$$

Hence, the likelihood equations are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\sigma\mu_i^3} \cdot \left(-\frac{1}{2}\mu^3\right) = -\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, \ldots, p,$$

which is equivalent to

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, \ldots, p.$$

d

As $a(\phi) = -2\phi = \phi/\omega_i$ with $\omega_i = -\frac{1}{2}$, the deviance can be expressed by the formula

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \omega_i \left(y_i \left(\tilde{\theta}_i - \widehat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right),$$

where $\widehat{\theta}_i = \frac{1}{\widehat{\mu}_i^2}$, $b(\widehat{\theta}_i) = \frac{2}{\widehat{\mu}_i}$ and $y_i = b'(\tilde{\theta}_i) = \frac{1}{\sqrt{\tilde{\theta}_i}}$, so that $\tilde{\theta}_i = \frac{1}{y_i^2}$ and $b(\tilde{\theta}_i = 2\sqrt{\tilde{\theta}_i} = \frac{2}{y_i}$. Hence the deviance is given by

$$\begin{aligned} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) =& 2\sum_{i=1}^{n} \left(-\frac{1}{2}\right) \left(y_i \left(\frac{1}{y_i^2} - \frac{1}{\widehat{\mu}_i^2} \right) - \frac{2}{y_i} + \frac{2}{\widehat{\mu}_i} \right) \\ =& \sum_{i=1}^{n} \left(\frac{y_i}{\widehat{\mu}_i^2} - \frac{2}{\widehat{\mu}_i} + \frac{1}{y_i} \right). \end{aligned}$$

e

Since $Y_1, \ldots, Y_n$ are independent, the log-likelihood function is given by

(Continued on page 3.)

$$L(\boldsymbol{\mu}, \sigma; \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \mu_i, \sigma) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi y_i^3 \sigma) - \frac{1}{2\sigma} \frac{(y_i - \mu_i)^2}{y_i \mu_i^2} \right)$$
$$= -\frac{1}{2} \sum_{i=1}^{n} \log(2\pi y_i^3) - \frac{n}{2} \log(\sigma) - \frac{1}{2\sigma} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{y_i \mu_i^2}.$$

This gives

$$\frac{\partial L(\boldsymbol{\mu}, \sigma; \mathbf{y})}{\partial \sigma} = -\frac{n}{2\sigma} + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{y_i \mu_i^2},$$

so that the maximum likelihood estimate $\widehat{\phi}_{MLE} = \widehat{\sigma}_{MLE}$ is given by

$$-\frac{n}{2\widehat{\phi}_{MLE}} + \frac{1}{2\widehat{\phi}_{MLE}^2} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{y_i \widehat{\mu}_i^2} = 0,$$

where $\widehat{\mu}_i$ is the maximum likelihood estimate of μ_i. This gives

$$\widehat{\phi}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{y_i \widehat{\mu}_i^2} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\widehat{\mu}_i^2} - \frac{2}{\widehat{\mu}_i} + \frac{1}{y_i} \right) = \frac{1}{n} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}),$$

so that the corresponding estimator is $\widehat{\phi}_{MLE} = \frac{1}{n} D(\mathbf{Y}; \widehat{\boldsymbol{\mu}})$, $\widehat{\boldsymbol{\mu}}$ denoting the maximum likelihood estimator of $\boldsymbol{\mu}$. Further, as $D(\mathbf{Y}; \widehat{\boldsymbol{\mu}})/\phi \overset{approx}{\sim} \chi^2_{n-p}$, it has a mean approximatley equal to $n - p$, so that $\frac{1}{n-p}\mathrm{E}(D(\mathbf{Y}; \widehat{\boldsymbol{\mu}})$ has mean approximately equal to ϕ, as opposed to the maximum likelihood estimator $\widehat{\phi}_{MLE}$, which is biased.

Problem 2

a

The fitted model is a logistic regression model, i.e. a binomial GLM with logit link, specified by:

- $Y_1, \ldots, Y_N$ are independent with $Y_i \sim bin(\pi_i, 1)$, $i = 1, \ldots, N$
- the linear predictor is given by $\eta_i = \mathbf{x}_i \boldsymbol{\beta}$, $i = 1, \ldots, N$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i8})$ with $x_{i1} = 1$, $x_{i2}, \ldots, x_{i6}$ corresponding to `Sex`, `Age`, `Height`, `Weight` and `Surface` and x_{i7} and x_{i8} are dummy variables for the three levels of `Vision`, and finally $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_8)^T$
- the mean $\mu_i = \mathrm{E}(Y_i)$ is related to the linear predictor via the link function as $\eta_i = g(\mu_i) = logit(\mu_i) = \mu_i/(1 - \mu_i)$.

In a logistic regression model, e^{β_j} is the relative effect on the odds of one unit's increase in the j-th covariate when the other covariates remain the same. Hence, since the estimated beta coefficient of `Height` is -0.096, the (estimated) relative effect on the odds of having a low balance score of being 1cm higher, all other covariates remaining the same, is $e^{-0.096} \approx 0.91$.

(Continued on page 4.)

b

The full table of analysis of deviance is given by:

```
Analysis of Deviance Table

Model 1: ctsib.low ~ Sex + Height + Surface + Vision
Model 2: ctsib.low ~ Sex + Height + Weight + Surface + Vision
Model 3: ctsib.low ~ Sex + Age + Height + Weight + Surface + Vision
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       474     301.53
2       473     295.21  1   6.3131  0.01198 *
3       472     295.20  1   0.0108  0.91740
```

where `Resid. Dev` for Model 2 is the deviance $D(M_2)$ of Model 2, whereas `Deviance` is $D(M_1) - D(M_2) = 6.3131$, $D(M_1) = 301.51$ being the deviance of Model 1, so that $D(M_2) = 301.51 - 6.3131 = 295.21$, `Resid. Df` for Model 3 is given by $N - p_3 = 480 - 8 = 472$, as Model 3 har $p_3 = 8$ β-coefficients, and `Deviance` for Model 3 is given by $D(M_2) - D(M_3) = 295.21 - 295.20 = 0.01$.

The analysis of deviance table shows the results from likelihood ratio tests, first comparing Model 1, without both `Sex` and `Weight` (but all the other main effects) to Model 2, with `Weight`, but without `Sex`, and then comparing Model 2, with Model 3, containing both `Sex` and `Weight`. The low P-value from the former test suggest that Model 2 is to be preferred over Model 1, i.e. that `Weight` should be in the model, whereas the P-value from the latter is very high, indicating that Model 2 should be preferred over Model 3, so that `Sex` should be removed from the model. One should therefore choose Model 2.

c

The AIC values of the probit regression model is lower than that of the logistic one (though not very much lower), which indicates that the probit link provides a better fit to these data. As for the interpretation of the β-coefficient of `Height`, it is by no means straighforward in the probit model. It corresponds to an increase of 1cm of the height on the linear predictor, but that is linked to the risk of low balance score through the quantile function of the standard normal distribution, which is difficult to interpret. Hence, if interpretation of the parameters of the model is important, the logit link may be preferred, especially since the AIC for the model with that link is not that much higher.

Problem 3

a

With $Z \sim N(0, 1)$, we have

$$P(Z \leq \mathbf{x}_{ij}\boldsymbol{\beta} + u_i) = \Phi(\mathbf{x}_{ij}\boldsymbol{\beta} + u_i)$$

(Continued on page 5.)

Thus, since $\Phi^{-1}\left(P(Y_{ij}=1|u_i)\right)=\mathbf{x}_{ij}\boldsymbol{\beta}+u_i$,

$$P(Y_{ij}=1|u_i)=\Phi(\mathbf{x}_{ij}\boldsymbol{\beta}+u_i)=P(Z\leq\mathbf{x}_{ij}\boldsymbol{\beta}+u_i|u_i).$$

Further, since both Z and u_i follow a normal distribution, so must $Z-u_i$, which is a linear combination of the two. We have

$$\begin{aligned}\mathrm{E}(Z-u_i)=&\mathrm{E}(Z)-\mathrm{E}(u_i)=0-0=0\\ \mathrm{Var}(Z-u_i)\overset{ind.}{=}&\mathrm{Var}(Z)+(-1)^2\cdot\mathrm{Var}(u_i)=1+\sigma_u^2,\end{aligned}$$

so that $Z-u_i\sim N(0,1+\sigma_u^2)$. This means that

$$\begin{aligned}P(Y_{ij}=1|u_i)=\Phi(\mathbf{x}_{ij}\boldsymbol{\beta}+u_i)=P(Z\leq\mathbf{x}_{ij}\boldsymbol{\beta}+u_i)=&P(Z-u_i\leq\mathbf{x}_{ij}\boldsymbol{\beta})\\ =&P\left(\frac{Z-u_i}{\sqrt{1+\sigma_u^2}}\leq\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}\right)\\ =&\Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}\right).\end{aligned}$$

Hece, we get

$$\begin{aligned}\mu_{ij}=\int\mathrm{E}(Y_{ij}|u_i)f(u_i;\sigma_u^2)du_i=&\int P(Y_{ij}=1|u_i)f(u_i;\sigma_u^2)du_i\\ =&\int\Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}\right)f(u_i;\sigma_u^2)du_i\\ =&\Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}\right)\underbrace{\int f(u_i;\sigma_u^2)du_i}_{=1}\\ =&\Phi\left(\frac{\mathbf{x}_{ij}\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}\right).\end{aligned}$$

This means that the marginal model of the binomial GLMM with probit link has the same link as the binomial GLM with probit link, but the β-coefficients are all divided by $\sqrt{1+\sigma_u^2}$.

b

The observed responses in **Problem 2** are 12 repeated balance scores for 40 different i ndividuals, under different conditions, and thus covariates values. It is natural to assume that the balance scores for different individuals are independent, but not the balance scores for the same individual. This dependence is not taken into account in the models considered in **Problem 2**, as they all assume that the responses are all indipendent. A way to handle the dependence between the reponses for the same individual, is precisely to introduce a random intercept, that is individual specific, treating the individuals as clusters.

(Continued on page 6.)

From the R output, we see that $\widehat{\sigma}_u^2 = 2.242$, which we can use to compute $\widehat{\beta}_j/\sqrt{1+\widehat{\sigma}_u^2}$, $j = 1, \ldots, 7$, and it turns out that these are approximately equal to the estimated β-coefficients of the probit model form **Problem 2c**, which is in line with the result from **a**. Further, it we look at the P-values from the Wald tests for the β-coefficients, we see that several of the effects that were siginificant in the GLM, more specifically `Sex` and `Weight`, are no longer sigificant. Hence, ignoring the dependence between the responses may lead us to think that covariates have a significant effect on the risk of low balance score, when in fact they do not.

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to Generalized Linear Models
Day of examination:	Thursday 14th December 2023
Examination hours:	15.00 – 19.00

This problem set consists of 6 pages.

Appendices:	Formulas in STK3100/4100
Permitted aids:	Approved calculator

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

Problem 1

We assume that $V \sim \text{bin}(n, \pi)$, i.e. the random variable V is binomially distributed with n trials and probability π of success in each trial. The probability mass function (PMF) of $Y = V/n$ can be written

$$P(Y = y) = \binom{n}{ny} \pi^{ny}(1-\pi)^{n-ny}, \quad y = 0, \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n-1}{n}, 1 \qquad (1)$$

a

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

$$f(y; \theta, \phi) = \exp\left\{(\theta y - b(\theta))/a(\phi) + c(y, \phi)\right\}$$

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

b

Use the expressions for $a(\phi)$ and $b(\theta)$ to

(i) show that $\mu = E(Y) = \pi$

(ii) determine $\text{Var}(Y)$

Assume in the following that the random variables $V_1, \ldots, V_N$ are independent with $V_i \sim \text{bin}(n_i, \pi_i)$, $i = 1, \ldots, N$, and let $Y_i = V_i/N_i$. Let $x_1, \ldots, x_N$ denote known explanatory variable values. Consider a generalized linear model (GLM) for $Y_1, \ldots, Y_N$, with linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, and canonical link function $g(\mu_i) = g(\pi_i) = \eta_i$.

(Continued on page 2.)

c

(i) Show that the canonical link function is

$$g(\pi_i) = \log \frac{\pi_i}{1-\pi_i}.$$

(ii) What is such a GLM called?

(iii) Determine the likelihood equations for this GLM, expressed using π_i, $i = 1, \ldots, N$.

d

Solving the likelihood equations provides the estimator $\widehat{\boldsymbol{\beta}}$, which has the approximate distribution

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1}\right)$$

Determine $\boldsymbol{W}$ such that the i'th diagonal element w_i is expressed by π_i and n_i.

e

(i) Determine the deviance for this GLM.

(ii) Assuming that the data is ungrouped, i.e. that $n_i = 1$, $i = 1, \ldots, N$, explain what the deviance can be used for.

Problem 2

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year, according to World Health Organization. In this problem we will consider data from $N = 3815$ residents in a town in USA, from a study with the goal of finding risk factors for coronary heart disease (CHD). We will consider how the probability of experiencing CHD during a 10 year period depends on the following explanatory variables

- `male`: Gender of individual (binary; 1: male, 0: female)
- `age`: Age of individual (numerical; in years)
- `currentSmoker`: Whether or not the individual is a smoker (binary; 1: smoker; 0: non-smoker)
- `cigsPerDay`: Average number of cigarettes the individual smoked (numerical; per day)
- `totChol`: Total cholesterol level (numerical)
- `sysBP`: Systolic blood pressure (numerical)
- `glucose`: Glucose level (numerical)

The response variable (`TenYearCHD`) is a binary indicator of whether the individual experienced CHD during the 10 years of study ("1") or not ("0").

(Continued on page 3.)

a

Below is given output from fitting a model with all the main effects described above in R.

(i) Describe the model used, including all necessary assumptions, and expressed in terms of response variable Y_i and explanatory variables $x_{i1}, \ldots, x_{i7}$, $i = 1, \ldots, N$. Be careful to specify what each variable represents.

(ii) Give an interpretation of the estimate belonging to the explanatory variable `male`.

```
 > summary(hd.fit1)

Call:
glm(formula = TenYearCHD ~ male + age + currentSmoker + cigsPerDay +
    totChol + sysBP + glucose, family = binomial, data = hd.data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.182009   0.467910 -19.623  < 2e-16 ***
male           0.549614   0.103885   5.291 1.22e-07 ***
age            0.066969   0.006262  10.694  < 2e-16 ***
currentSmoker  0.047117   0.151228   0.312   0.7554
cigsPerDay     0.018511   0.006046   3.061   0.0022 **
totChol        0.002459   0.001060   2.320   0.0204 *
sysBP          0.016992   0.002098   8.101 5.46e-16 ***
glucose        0.007620   0.001647   4.626 3.72e-06 ***
---

    Null deviance: 3289.6  on 3814  degrees of freedom
Residual deviance: 2908.2  on 3807  degrees of freedom

AIC: 2924.2
```

b

On the next page is given more output from R.

(i) Give two reasons from this output that indicates that `currentSmoker` should be dropped from the model.

(ii) Discuss why smoking status does not seem to be significant in this model.

(Continued on page 4.)

```
> drop1(hd.fit1,test="LRT")
Single term deletions

Model:
TenYearCHD ~ male + age + currentSmoker + cigsPerDay + totChol +
    sysBP + glucose
             Df Deviance    AIC     LRT  Pr(>Chi)
<none>            2908.2 2924.2
male          1   2936.4 2950.4  28.159 1.118e-07 ***
age           1   3027.3 3041.3 119.100 < 2.2e-16 ***
currentSmoker 1   2908.3 2922.3   0.097  0.755622
cigsPerDay    1   2917.5 2931.5   9.257  0.002346 **
totChol       1   2913.6 2927.6   5.312  0.021179 *
sysBP         1   2974.0 2988.0  65.760 5.093e-16 ***
glucose       1   2929.9 2943.9  21.643 3.284e-06 ***
```

c

We will now consider fits of three different models with all main effects except `currentSmoker`, two with interaction terms. A summary of these fits in the form of an analysis of variance table is given below. Some of the numbers have been replaced by question marks.

(i) Determine the numbers that have been replaced by question marks.

(ii) Which of the models has the best fit? Give an explanation.

```
Analysis of Deviance Table

Model 1: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose
Model 2: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose +
    totChol:glucose
Model 3: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose +
    totChol:glucose + totChol:sysBP
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3808     2908.3
2      ?        2904.8  1      ?    0.05926 .
3      3806       ?     1   0.9741  0.32366
```

Problem 3

In this problem we will consider response variables that represent counts, which are allowed to be correlated within groups. Let Y_{ij} denote response for subject i, $j = 1, \ldots, d$ in group j, $i = 1, \ldots, n$, and x_{ij} be a known explanatory variable value. A mixed Poisson generalized mixed model (GLMM) with log-link and random intercept $u_i \sim N(0, \sigma_u^2)$, $i = 1, \ldots, n$ is then given by that the Y_{ij}'s conditional on u_i are Poisson-distributed with conditional mean $E(Y_{ij} \mid u_i)$, with

$$\log\left(E(Y_{ij} \mid u_i)\right) = \beta_0 + \beta_1 x_{ij} + u_i$$

(Continued on page 5.)

a

(i) Show that the marginal (unconditional) mean $\mu_{ij} = E(Y_{ij})$ can be expressed as

$$E(Y_{ij}) = \exp(\beta_0 + \beta_1 x_{ij})\, E(\exp(u_i))$$

(ii) Determine $E(\exp(u_i))$ as a function of σ_u^2. Hint: Use that the moment generating function for the $N(0, \sigma_u^2)$ is $M(t) = \exp\left(\sigma_u^2 t^2/2\right)$.

(iii) Comment on the relationship between the fixed effects (intercept and effect of the explanatory variable) of the log-link Poisson GLMM and the marginal model $E(Y_{ij}) = \exp(\beta_0 + \beta_1 x_{ij})$.

b

In this part we will consider a dataset where the response variable counts the number of awards each of 200 high school students have recieved. The students come from 20 different schools, and the responses are assumed to be correlated within a school, but independent between different schools. The explanatory variable we will consider is the gender of the students. Below (continues on the next page) you see output from fitting two different models to this data; a GLMM and a marginal model fitted by generalized estimating equations (GEE).

```
Generalized linear mixed model fit by maximum likelihood
  (Adaptive Gauss-Hermite Quadrature, nAGQ = 100) [glmerMod]
 Family: poisson  ( log )
Formula: awards ~ 1 + female + (1 | cid)
   Data: award.data

     AIC      BIC   logLik deviance df.resid
   221.1    231.0   -107.6    215.1      197

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5312 -0.5919 -0.3304  0.2047  2.8806

Random effects:
 Groups Name        Variance Std.Dev.
 cid    (Intercept) 1.431    1.196
Number of obs: 200, groups:  cid, 20

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.2229     0.2975  -0.749  0.45370
femalefemale   0.3632     0.1193   3.044  0.00234 **

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
```

(Continued on page 6.)

```
 Link:                       Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Exchangeable

Call:
gee(formula = awards ~ 1 + female, id = cid, data = award.data,
    family = poisson, corstr = "exchangeable")

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-1.9440514 -1.3583181 -0.3583181  0.6416819  5.6416819


Coefficients:
              Estimate Naive S.E.  Naive z Robust S.E.
(Intercept)  0.3062472  0.2239515 1.367472   0.2310288
femalefemale    ?       0.1107031 3.238633   0.1228721
             Robust z
(Intercept)  1.325580
femalefemale 2.917886

Estimated Scale Parameter:  1.957069
```

(i) In the GEE fit of the marginal model, there is a question mark instead of the estimated coefficient for gender. Determine the missing number.

(ii) In the GEE fit of the marginal model, you see two columns with standard errors for the estimated coefficients; called "Naive S.E." and the "Robust S.E.". Explain briefly the difference between these two.

(iii) What is the method behind finding the numbers in the column "Robust S.E." called?

APPENDIX: Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \ldots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \ldots, \eta_n$ are given by $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and

the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right), \quad \text{approximately}$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \widehat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ for $i = 1, \ldots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots \epsilon_{id})^{\mathrm{T}} \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \,|\, \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to Generalized Linear Models
Day of examination:	Thursday 8th December 2022
Examination hours:	15.00 – 19.00
This problem set consists of 6 pages.	
Appendices:	Formulas in STK3100/4100
Permitted aids:	Approved calculator

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

Problem 1

We assume that $Y \sim \text{Poisson}(\mu)$, i.e. the random variable Y is Poisson distributed with parameter μ, and hence has probability mass function (PMF)

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

a

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

$$f(y; \theta, \phi) = \exp\{(\theta y - b(\theta))/a(\phi) + c(y, \phi)\} \tag{2}$$

and determine θ, $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$.

We then assume that $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \text{Poisson}(\mu_i)$, and hence $E(Y_i) = \mu_i$, $i = 1, \ldots, n$.

b

Write down the definition of a generalized linear model (GLM) for $Y_1, \ldots, Y_n$ with associated covariates $x_{ij}, i = 1, \ldots n, j = 1, \ldots, p$, with $x_{i1} = 1$ to represent the intercept. Use the canonical link function.

c

Assume that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an observed value of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. Derive the expression for the log-likelihood function $L(\boldsymbol{\mu}; \mathbf{y})$. Explain briefly what the saturated model is, and express the maximum of the log-likelihood $L(\boldsymbol{\mu}; \mathbf{y})$ for the saturated model.

(Continued on page 2.)

d

Find an expression for the deviance $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ for a Poisson GLM, and explain how it can be used to compare two different models.

Problem 2

In this problem we assume that Y comes from a negative binomial distribution. Here we let the pmf of the negative binomial distribution take the form

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

We will assume that $k > 0$ is a given constant, and consider the random variable $Y^* = Y/k$. Then $P(Y^* = y^*) = P(Y = ky^*)$ for $y^* = 0, \frac{1}{k}, \frac{2}{k}, \ldots$, so Y^* has pmf

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

a

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

b

Find the mean and variance of Y^* using the expressions for $b(\theta)$ and $a(\phi)$. Use these results to show that $E(Y) = \mu$ and determine $\text{var}(Y)$.

c

Compare the relationship between the mean and variance for negative-binomial distribution to the relationship between the mean and variance of the Poisson distribution, and comment on when the Poisson is a good model and when you need the negative-binomial. What does overdispersion mean?

Problem 3

In this problem we will consider data collected in Arizona in 1991 on patients entering the hospital to receive one of two standard cardiovascular procedures: Coronary Artery Bypass Graft (CABG) and Percutaneous Transluminal Coronary Angioplasty (PTCA). CABG involves taking a blood vessel from another part of the body and attaching it to the coronary artery above and below the narrowed area or blockage, so the the diseased sections are bypassed. PTCA, is a method of placing a balloon in a blocked coronary artery to open it to blood flow. The data set contains data on 3589 patients, and the response variable is length of hospital stay (los). For modeling this, we will consider the following covariates

(Continued on page 3.)

- `procedure`: Type of procedure (1: CABG, 0: PTCA)
- `sex`: Sex of patient (1: male, 0: female)
- `admit`: Type of admission (1: Urgent/Emergency; 0: elective/pre-planned)
- `age75`: Age group of patient (1: Age>75, 0: Age<=75)

a

We first fit a model with only main effects. The result of this analysis is given below. Describe the model used in this analysis, including all assumptions.

```
> fit1 <- glm(los ~ procedure + sex + admit+ age75, family=poisson)
> summary(fit1)

Call:
glm(formula = los ~ procedure + sex + admit + age75, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.45599    0.01585  91.874   <2e-16
procedure    0.96034    0.01218  78.836   <2e-16
sex         -0.12393    0.01181 -10.492   <2e-16
admit        0.32659    0.01212  26.939   <2e-16
age75        0.12222    0.01245   9.817   <2e-16
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16265.0  on 3588  degrees of freedom
Residual deviance:  8874.1  on 3584  degrees of freedom
AIC: 22390
```

b

We then fit a model with interaction between `procedure` and `admit`. In the following you find the results from this fit, followed by an analysis of deviance table for the two fits. Explain why the model with interactions is to be preferred over the model with only main effects. Describe the effects of `procedure` and `admit` on the estimated mean length of stay.

```
> fit2 = glm(los ~ procedure + sex + admit+ age75 + procedure*admit, family=poisson)
> summary(fit2)

Call:
glm(formula = los ~ procedure + sex + admit + age75 + procedure *
    admit, family = poisson)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.23851    0.02302  53.790   <2e-16
```

(Continued on page 4.)

```
procedure        1.24765    0.02417  51.613   <2e-16
sex             -0.12488    0.01182 -10.568   <2e-16
admit            0.61606    0.02426  25.395   <2e-16
age75            0.12314    0.01245   9.889   <2e-16
procedure:admit -0.39658    0.02803 -14.149   <2e-16
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16265.0  on 3588  degrees of freedom
Residual deviance:  8666.6  on 3583  degrees of freedom
AIC: 22184

> anova(fit1,fit2,test="LRT")
Analysis of Deviance Table

Model 1: los ~ procedure + sex + admit + age75
Model 2: los ~ procedure + sex + admit + age75 + procedure * admit
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      3584     8874.1
2      3583     8666.6  1   207.54 < 2.2e-16
```

c

To address the possible issue of overdispersion, we fit a negative-binomial model. The result of the analysis is given below. What do the AIC values for this model and the model in b tell you about which model to prefer? In the lectures and in the text book we have seen the parametrization $\gamma = 1/k$. Below you also find a transcript of the calculation of a test statistic and an accompanying p-value for the hypotheses test

$$H_0: \quad \gamma = 0 \qquad H_a: \quad \gamma > 0$$

What is the conclusion from this test? Does it support your conclusion from the AIC values?

```
> fit3 = glm.nb(los ~ procedure + sex + admit + age75 + procedure*admit)
> summary(fit3)

Call:
glm.nb(formula = los ~ procedure + sex + admit + age75 + procedure *
    admit, init.theta = 6.521921816, link = log)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.24084    0.02958  41.943  < 2e-16
procedure        1.24900    0.03218  38.813  < 2e-16
sex             -0.12745    0.01885  -6.761 1.37e-11
admit            0.61482    0.03073  20.009  < 2e-16
age75            0.12198    0.01999   6.101 1.05e-09
procedure:admit -0.39742    0.03894 -10.205  < 2e-16
(Dispersion parameter for Negative Binomial(6.5219) family taken to be 1)
```

(Continued on page 5.)

```
    Null deviance: 6869.3  on 3588  degrees of freedom
Residual deviance: 3504.0  on 3583  degrees of freedom
AIC: 19857

              Theta:  6.522
          Std. Err.:  0.268

 2 x log-likelihood:  -19843.162
> test.statistic=-as.numeric(2*(logLik(fit2)-logLik(fit3)))
> p.value=(1-pchisq(test.statistic,1))/2
> print(p.value)
[1] 0
```

Problem 4

The negative-binomial addresses the issue of overdispersed count data that has a variance greater than the mean. A more general approach which can be used also for other types of overdispersed data is called the quasi-likelihood approach for overdispersion. It is based on $\text{var}(Y_i) = \phi v^*(\mu_i)$, where ϕ is an overdispersion parameter and $v^*(\mu_i)$ is the function specifying how the variances from the "standard" GLM depend on the expected values $\mu_i = E[Y_i]$. For the Poisson distribution $v^*(\mu_i) = \mu_i$. A transcript of an analysis with the quasi-likelihood variance inflation approach to the data from Problem 3, with the same covariates as in Problem 3b, can be seen below.

a

Compare the estimated β_j's and their estimated standard errors to the ones for the Poisson GLM fit `fit2` and comment.

b

Explain in general terms the approach, based on the quasi-likelihood equations for the estimation. Hint: Replace $\text{var}(Y_i)$ by $v(\mu_i)$ in the likelihood equations for a GLM that you find in the appendix.

```
> fit.quasipois=glm(los ~ procedure + sex + admit+ age75+ procedure*admit,
+                   family=quasi(link="log",variance="mu"))
> summary(fit.quasipois)

Call:
glm(formula = los ~ procedure + sex + admit + age75 + procedure *
    admit, family = quasi(link = "log", variance = "mu"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1102  -1.1806  -0.5060   0.5216  12.8660
```

(Continued on page 6.)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.23851    0.04087  30.301  < 2e-16
procedure        1.24765    0.04291  29.074  < 2e-16
sex             -0.12488    0.02098  -5.953 2.88e-09
admit            0.61606    0.04307  14.305  < 2e-16
age75            0.12314    0.02210   5.571 2.72e-08
procedure:admit -0.39658    0.04976  -7.970 2.11e-15

(Dispersion parameter for quasi family taken to be 3.151389)

    Null deviance: 16265.0  on 3588  degrees of freedom
Residual deviance:  8666.6  on 3583  degrees of freedom
```

APPENDIX: Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \ldots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \ldots, \eta_n$ are given by $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and

the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right), \quad \text{approximately}$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \widehat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ for $i = 1, \ldots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots \epsilon_{id})^{\mathrm{T}} \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \,|\, \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to Generalized Linear Models
Day of examination:	Monday 6th December 2021
Examination hours:	15.00 – 19.00
This problem set consists of 5 pages.	
Appendices:	Formulas in STK3100/4100
Permitted aids:	Approved calculator

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

Problem 1

A distribution with probability mass function (pmf) or probability density function (pdf)

$$f(y;\theta,\phi) = \exp\left\{\left(\theta y - b(\theta)\right)/a(\phi) + c(y,\phi)\right\} \tag{1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are functions, and θ and ϕ parameters, belongs to the *exponential dispersion family* of distributions.

a

Show that if the random variable Y has a distribution from the exponential dispersion family with parameters θ and ϕ, then $E[Y] = b'(\theta)$ and $\text{Var}[Y] = b''(\theta)a(\phi)$.

Hint: Use that we have from general likelihood results that

$$E\left[\frac{\partial \log f(Y;\theta,\phi)}{\partial\theta}\right] = 0$$

$$-E\left[\frac{\partial^2 \log f(Y;\theta,\phi)}{\partial\theta^2}\right] = E\left[\left(\frac{\partial \log f(Y;\theta,\phi)}{\partial\theta}\right)^2\right].$$

and in addition that $\text{Var}[Y] = E\left[(Y - E[Y])^2\right]$.

In the following we assume that the random variable Y is Poisson distributed with pmf

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

b

Show that the Poisson distribution belongs to the exponential dispersion family of distributions, that is show that (2) can be written in the form of (1). Determine θ, $a(\phi)$, $b(\theta)$ and $c(y,\phi)$.

(Continued on page 2.)

c

Show that $E[Y] = \text{Var}[Y] = \mu$.

Assume now that $Y_1, Y_2, \ldots, Y_n$ are independent and that Y_i is Poisson distributed with mean μ_i, $i = 1, \ldots, n$. Given the values x_i, $i = 1, \ldots, n$, of an explanatory variable, we assume a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_n$, with the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$ linked to the mean through the canonical link function $g(\mu_i) = \log(\mu_i) = \eta_i$.

d

Derive an expression for the log-likelihood function $L(\beta_0, \beta_1)$, and show that the maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions to the following equations

$$\sum_{i=1}^{n}(Y_i - \mu_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n}(Y_i - \mu_i)x_i = 0$$

Problem 2

In this problem we will consider data collected for a study analysing the probability of objections agains patents granted by the European patent office. The data set concerns 2702 patents from the sector of semiconductor/computer industry. On each of the 2702 patents the following variables are recorded

- `opp`: Patent opposition (1=yes; 0=no)
- `year`: Grant year
- `ncit`: Number of citations for the patent
- `ustwin`: US twin patent exists (1=yes; 0=no)
- `patus`: Patent holder from the US (1=yes; 0=no)
- `patgsgr`: Patent holder from Germany, Switzerland or Great Britain (1=yes; 0=no)
- `ncountry`: Number of designated countries for the patent

We will investigate how the probability of having an objection against a patent depends on grant year, number of citations for the patent, whether a US twin patent exists, whether the patent holder is from the US, whether the patent holder is from Germany, Switzerland or Great Britain, and the number of designated countries for the patent.

a

On the next page you find output from R from a fit with only main effects. Describe the model behind this fit, including necessary assumptions. Give an interpretation of the estimate for `patus`.

(Continued on page 3.)

```
 Call:
glm(formula = opp ~ year + ncit + ustwin + patus + patgsgr +
    ncountry, family = binomial(link = logit), data = patents)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7594  -0.8181  -0.6328   1.1397   2.2151

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 185.25738   21.85850   8.475  < 2e-16 ***
year         -0.09373    0.01098  -8.537  < 2e-16 ***
ncit          0.12128    0.02214   5.477 4.33e-08 ***
ustwin       -0.41085    0.09959  -4.126 3.70e-05 ***
patus        -0.43685    0.10998  -3.972 7.12e-05 ***
patgsgr       0.18723    0.11691   1.602    0.109
ncountry      0.10302    0.01481   6.957 3.48e-12 ***
---
Null deviance: 3214.5  on 2701  degrees of freedom
Residual deviance: 2996.8  on 2695  degrees of freedom
AIC: 3010.8
```

b

We have also fitted models with interactions. A summary of the fits can be seen in the R output of an analysis of variance table below. Some of the entries in the table are missing and have been replaced by question marks. Give the missing numbers, along with an explanation of how you found them. Which of the models fit the data best, and why?

```
     > anova(fit1,fit2,fit3,fit4,test="LRT")
Analysis of Deviance Table

Model 1: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry
Model 2: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry
     + year:patus
Model 3: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry
    + year:patus + patus:ncountry
Model 4: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry
    + year:patus + patus:ncountry + patgsgr:ncountry
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2695     2996.8
2      2694     2981.7  ?      ?    0.0001011 ***
3        ?      2979.2  1  2.5449   0.1106478
4      2692       ?     1  2.5899   0.1075493
```

(Continued on page 4.)

c

The table below shows the AIC values for a binomial GLM-model with three different link functions, and the same linear predictor as for the best model in b. Explain what AIC is. Argue why it is sensible to use this as a criteria to compare the different models arising from the different link-functions in this situation, instead of a test such as the likelihood ratio test. Based on these values, which link function should be chosen?

Link function	AIC
logit	2997.724
probit	2996.969
cloglog	3000.778

Problem 3

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{id})^T$, $i = 1, \ldots, n$, be independent vectors that correspond to observations from each of n groups. Given model matrices $\mathbf{X}_i$ (of dimension $d \times p$, with vector $\mathbf{x}_{ij}$ for observation j from group i in row j) and $\mathbf{Z}_i$ (of dimension $d \times q$ with vector $\mathbf{z}_{ij}$ for observation j from group i in row j) , we assume a normal linear mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})^T \sim N(\mathbf{0}, \mathbf{R})$ is independent of $\mathbf{u}_i$.

Now with

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_n \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix},$$

$$\boldsymbol{\Sigma}_\mathbf{u} = \begin{pmatrix} \boldsymbol{\Sigma}_u & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_u & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_u \end{pmatrix}, \quad \mathbf{R}_\varepsilon = \begin{pmatrix} \mathbf{R} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R} \end{pmatrix}$$

we may write $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$. Furthermore, we have that

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T + \mathbf{R}_\varepsilon$. We assume throughout this problem that that $\mathbf{V}$ is known.

a

Given observed data $\mathbf{Y} = \mathbf{y}$, show that the maximum likelihood estimate of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Hint: Use the matrix derivative results

$$\partial(\boldsymbol{a}^T\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = \boldsymbol{a} \quad \text{and} \quad \partial(\boldsymbol{\gamma}^T\boldsymbol{A}\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{\gamma}$$

(Continued on page 5.)

b

Show that

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{u}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V} & \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u}\\ \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T & \boldsymbol{\Sigma}_\mathbf{u}\end{pmatrix}\right]$$

and that

$$E[\boldsymbol{u}|\mathbf{Y}=\mathbf{y}] = \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

Given $\mathbf{Y}=\mathbf{y}$ and $\tilde{\boldsymbol{\beta}}$, what is a sensible prediction for the random effects?

c

With $\mathbf{Y}_i$, $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$, $\boldsymbol{\beta}$ and $\mathbf{u}_i$ defined as before, the *generalized* linear mixed model (GLMM) for $\mathbf{Y}_i$ has the form

$$g\left(E[Y_{ij}|\mathbf{u}_i]\right) = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i, \quad i=1,\ldots,n, \quad j=1,\ldots,d.$$

Still assuming that g is the identity link function, find the marginal expected value $\mu_{ij} = E[Y_{ij}]$. Comment on the link function for this marginal model implied by the GLMM. Is this relationship a general result for all link functions?

END

STK3100/STK4100 - Introduction to Generalized Linear Models

Additional information

- In **Problem 3**, $\mathbf{\Sigma_u}$ and $\mathbf{R}_\varepsilon$ can be assumed known.
- Hint for **Problem 3 c**: $E[Y] = E[E[Y|X]]$

APPENDIX: Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \ldots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \ldots, \eta_n$ are given by $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and

the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right), \quad \text{approximately}$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \widehat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ for $i = 1, \ldots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots \epsilon_{id})^{\mathrm{T}} \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \,|\, \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100 / STK4100 — Introduction to generalized linear models.

Day of examination: Wednesday December 2nd 2020

Examination hours: 15.00 – 19.30 (including 30 minutes for Inspera delivery).

This problem set consists of 4 pages.

Appendices: None

Permitted aids: All resources

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

Problem 1

a) Assume that Y is Poisson-distributed with $\mathrm{P}(Y = y; \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$ for $y = 0, 1, \ldots$

Show that the Poisson-distribution can be written on the natural exponential family form $\mathrm{P}(Y = y; \lambda) = \exp(\theta y - b(\theta) + c(y))$. In particular identify the canonical parameter θ, the cumulant function $b(\theta)$ and $c(y)$.

Use $b(\theta)$ to derive expressions for $\mu = \mathrm{E}[Y]$ and $\mathrm{var}[Y]$.

b) Show that the truncated Poisson distribution for $Y \mid Y > 0$ has probability mass function $\frac{\lambda^y}{y!} \exp(-\lambda)/(1 - \exp(-\lambda))$ for $y = 1, 2, \ldots$.

Verify that also this distribution can be written on the natural exponential family form $\exp(\theta y - b(\theta) + c(y))$ and identify θ, $b(\theta)$ and $c(y)$ in this case.

c) As a generalization of this assume that Y has density or probability mass function $f(y; \gamma) = \exp(\gamma y - b_0(\gamma) + c_0(y))$ over a set S of permissible values of y and assume that B is a subset of S with $\mathrm{P}(Y \in B) > 0$ for all possible γ.

Show that in general $Y \mid Y \in B$ has a distribution on the natural exponential family form with density or probability mass function $f_B(y; \theta) = \exp(\theta y - b(\theta) + c(y))$ for $y \in B$. In particular identify θ and give an expression for $b(\theta)$.

(Continued on page 2.)

Problem 2

The data for this problem stems from an investigation of whether a health reform in Germany in 1997 led to reduced number of doctoral visits among women aged 20-40 years. Some individuals were interviewed in 1996 (before the reform) and others in 1998 (after the reform). The women reported the number of doctoral visits the last year. In this problem this response has been dichotomized into unfrequent and frequent visitors (somewhat arbitrarily) defined as $Y_i = 0$ if the number of visits were below 7 and $Y = 1$ if the number of visits were 7 or more. These responses Y_i are analysed with logistic regression for x_{i1} indicating interview before or after the reform (0=before, 1=after), x_{i2} = indicator of poor or very poor health (1=yes, 0=no, loosely referred to as "bad health") and other explanatory variables.

a) In a first model only x_{i1} and x_{i2}, denoted as `reform` and `badh` in the `R`-output below, were included in the logistic regression with binary outcomes defined in `R` as `I(numvisit>6)`.

Give interpretations of $\exp(\hat{\beta}_j)$, where $\hat{\beta}_j$ are the estimates of the regression coefficient for x_{ij} for $j = 1$ and 2,

(i) in general

(ii) as an approximation valid when $\mathrm{P}(Y_i = 1)$ are all small (for these data the overall proportion $Y_i = 1$ was 0.086).

```
> fit=glm(I(numvisit>6)~badh+reform,family=binomial,data=drv)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7180     0.1230 -22.103   <2e-16
badh          2.1995     0.1668  13.189   <2e-16
reform       -0.3382     0.1619  -2.089   0.0367
```

b) Explain why $\exp(\hat{\beta}_j \pm 1.96 se_j)$, where se_j are the standard errors of the $\hat{\beta}_j$, are approximate 95% confidence intervals for $\exp(\beta_j)$.

Calculate these confidence intervals for $\exp(\beta_1)$ and $\exp(\beta_2)$ and use the intervals to determine conclusions to hypothesis tests for $\mathrm{H}_{0j} : \beta_j = 0$ versus $\mathrm{H}_{0j} : \beta_j \neq 0$ with a 5% significance level.

Compare the conclusions of the tests with the Wald z-values and p-values from the `R`-output.

(Continued on page 3.)

c) Below you find R-output from an extended model where also the explanatory variables education (categorized to three levels <10, 10-12 and > 12 years, `educat`) and income (categorized to three levels, `inccat`) as well as interactions between "bad" health and reform, "bad" health and income and education and income are included. The output is a deviance table with certain values replaced with question marks.

Explain what deviances are and what deviance table are used for.

Fill in correct values where 4 numbers are replaced by question marks.

```
> fit=glm(I(numvisit>6)~badh+reform+educat+inccat+badh:reform
       +badh:inccat+educat:inccat ,family=binomial,data=drv)
> anova(fit,test="Chi")

Analysis of Deviance Table
Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2226     1303.4
badh           1  158.613      2225     1144.8 < 2.2e-16
reform         1        ?      2224     1140.4  0.035849
educat         2    2.339       ?         ?     0.310536
inccat         2    8.641      2220     1129.4  0.013292
badh:reform    1    1.458      2219     1127.9  0.227285
badh:inccat    2    0.851      2217     1127.1  0.653313
educat:inccat  ?   13.689      2213     1113.4  0.008357
```

Problem 3

The log-normal distributions are not an exponential dispersion family. We will in this problem see how we may still use results or theory for GLM to fit regression models for responses that are log-normal.

a) Assume that Y is a log-normal random variable and so is defined by that $V = \log(Y)$ is a normal distributed random variable with mean $\mathrm{E}[V] = \gamma$ and $\mathrm{var}[V] = \sigma^2$.

Show that with $\mu = \mathrm{E}[Y]$ one can express $\mathrm{var}[Y] = \phi\mu^2$ where the dispersion parameter $\phi = \exp(\sigma^2) - 1$.

Hint: You can use that the moment-generating function of $V \sim \mathrm{N}(\gamma, \sigma^2)$ can be expressed as $M_V(t) = \mathrm{E}[\exp(tV)] = \exp(\gamma t + \frac{1}{2}t^2\sigma^2)$.

(Continued on page 4.)

b) Assume $Y_1, \ldots, Y_n$ are independent and log-normally distributed with expected values $\mu_i = \mathrm{E}[Y_i] = \exp(\alpha + \beta x_i + \frac{1}{2}\sigma^2)$ where the x_i are known explanatory variables with the same variance σ^2 of $V_i = \log(Y_i)$.

Demonstrate how simple linear regression for the $V_i = \log(Y_i)$ can be used to obtain estimates of α, β and ϕ.

c) If on the other side $\mu_i = \alpha + \beta x_i$ such a simple linear regression technique can not be applied to solve the estimation problem.

However one can use a technique from GLM-theory based on the relationship $\mathrm{var}[Y_i] = \phi\nu^*(\mu_i)$. Then $\nu^*(\mu_i)$ is a function specifying how the variances depend on the expected values $\mu_i = \mathrm{E}[Y_i]$. For the log-normal distribution we have from question a) that $\nu^*(\mu_i) = \mu_i^2$.

Explain in general terms this approach. It can be useful to state appropriate score equations for the estimation.

Problem 4

A model for binary matched pair data (Y_{i1}, Y_{i2}) with explanatory variables x_{ij} and random intercept u_i is written as

$$\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | x_{ij}, u_i)) = \beta_0 + \beta_1 x_{ij} + u_i$$

where $\mathrm{logit}(\pi) = \log(\pi/(1-\pi))$.
We assume that the u_i has a density $f(u; \sigma^2)$, typically from a $\mathrm{N}(0, \sigma_u^2)$-distribution, and that conditionally on u_i we have Y_{i1} and Y_{i2} independent. We also assume that the u_i's are independent so that the pairs $(Y_{i1}, Y_{i2}); i = 1, \ldots, n$ are independent.

a) Present an expression for the marginal probability $\mathrm{P}(Y_{ij} = 1 | x_{ij})$.

Argue that for each pair, Y_{i1} and Y_{i2} are marginally dependent.

Set up an expression for the marginal likelihood $l(\beta_0, \beta_1, \sigma_u^2)$.

b) Alternatively we may consider the u_i's as fixed effects and estimate β_1 by conditioning on $Y_{i1} + Y_{i2}$. Show that

$$\mathrm{P}(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 1) = \frac{\exp(\beta_1(x_{i1} - x_{i2}))}{1 + \exp(\beta_1(x_{i1} - x_{i2}))}$$

and argue that $\mathrm{P}(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 2) = 1 = \mathrm{P}(Y_{i1} = 0 | Y_{i1} + Y_{i2} = 0)$. (In these expression the conditioning on the observed numbers x_{ij} has been supressed from the notation).

Explain based on this how logistic regression can be set up to obtain an estimate of β_1.

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100 / STK4100 — Introduction to generalized linear models.
Day of examination:	Wednesday December 18th 2019
Examination hours:	9.00 – 13.00.
This problem set consists of 5 pages.	
Appendices:	Formulas in STK3100 / STK4100
Permitted aids:	Approved calculator

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

Problem 1

a) Assume that $Y_i, i = 1, \ldots, n$ are independent binary responses with $\pi_i = \mathrm{P}(Y_i = 1) = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$. Here x_i is an explanatory variable and α and β regression parameters.

Show that $\exp(\beta)$ can be interpreted as an odds-ratio.

Give an approximate interpretation of $\exp(\beta)$ valid when the π_i's are small.

b) In a case-control study, where typically the π_i's are small, one will only collect data on the explanatory variable for a subset of the observations but will oversample the observations with $Y_i = 1$. Thus, one collects x_i for observation i if a sampling indicator equals 1, that is $Z_i = 1$, where

$$\rho_1 = \mathrm{P}(Z_i = 1|Y_i = 1) \quad \text{and} \quad \rho_0 = \mathrm{P}(Z_i = 1|Y_i = 0),$$

allowing for different sampling fractions ρ_1 and ρ_0 for cases ($Y_i = 1$) and controls ($Y_i = 0$). Note that ρ_j can not depend on x_i.

Show that

$$\mathrm{P}(Y_i = 1|Z_i = 1) = \frac{\exp(\alpha^* + \beta x_i)}{1 + \exp(\alpha^* + \beta x_i)}$$

where $\alpha^* = \alpha + \log(\rho_1/\rho_0)$.

What does this result mean in practice for analyzing case-control data?

Hint: First show that $P(Z_i = 1) = \rho_1\pi_i + \rho_0(1 - \pi_i)$ or use Bayes theorem directly.

(Continued on page 2.)

Problem 2

a) The density for the gamma distribution can be expressed as

$f(y;\mu,k) = (k/\mu)^k y^{k-1}\exp(-(k/\mu)y)/\Gamma(k)$ for $y > 0$.

Show that the gamma distribution density can be rewritten on the exponential dispersion family form $\exp((\theta y - b(\theta))/\phi + c(y,\phi))$ with $\theta = -1/\mu$, $b(\theta) = -\log(-\theta)$ and $\phi = 1/k$.

Verify that when $Y \sim f(y;\mu,k)$ then $\mathrm{E}[Y] = \mu$ and $\mathrm{var}[Y] = \phi\mu^2$.

b) Write down the definition of a generalized linear model with gamma distributed responses $Y_i, i = 1,\ldots,n$.

Verify that the likelihood-equations for such a model can be written as

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi\mu_i^2} x_{ij} \frac{\partial\mu_i}{\partial\eta_i} = 0 \quad \text{for } j = 1,\ldots,p$$

with observations y_i of Y_i, x_{ij} is explanatory variable $j = 1,\ldots,p$ and η_i the linear predictor for observation i.

c) The estimators of the regression coefficients determined by solving the equations in question b) are valid also when the Y_i's are not gamma distributed as long as the expected values μ_i and the variance structure $\mathrm{var}[Y_i] = \phi\mu_i^2$ are correctly specified.

Give a brief explanation for why this is true.

Suggest an estimator for the dispersion term ϕ which is valid both when the Y_i's are gamma distributed and when only the expectation and variance structure are correctly specified.

d) Prices of $n = 100$ apartments sold in Oslo in the year 2000 were collected along with explanatory variables x_1 = area in m^2 (in R-output `size`), x_2 = no. of rooms in the apartment (`rooms`), x_3 = indicator if the apartment has a balcony or not (`balcony`), x_4 = monthly expenses or rent in NOK (`rent`) and x_5 = location of apartment in west/east direction measured in km (low numbers means in west, high in east of Oslo) (`x`). Results from analyzing the prices with a gamma GLM with an identity link are given on the next page.

Give a description of how the explanatory variables affect the price of the apartments.

Identify which explanatory variables significantly influence the price.

Find the estimated apartment price when $x_1 = 70\ \mathrm{m}^2$, $x_2 = 2$ rooms, $x_3 = 0$, i.e. no balcony, $x_4 = 1000$ NOK and $x_5 = 2$ km to the east.

How would you determine the uncertainty of this estimate (an exact numerical answer is not possible with the given information).

(Continued on page 3.)

```
Call:
glm(formula = price ~ size + rooms + balcony + rent + x,
       family = Gamma(link = identity))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 526.64340   48.15544  10.936  < 2e-16 ***
size         18.39589    1.31699  13.968  < 2e-16 ***
rooms        25.77173   30.88332   0.834  0.40612
balcony      81.67536   30.14617   2.709  0.00801 **
rent         -0.12745    0.01368  -9.318 5.19e-15 ***
x           -93.28967   10.29277  -9.064 1.80e-14 ***
---
(Dispersion parameter for Gamma family taken to be 0.0167957)

    Null deviance: 15.5898  on 99  degrees of freedom
Residual deviance:  1.6817  on 94  degrees of freedom
```

e) Since an identity link has been used one may also consider analyzing the apartment prices with linear regression. Below you find results from such an analysis with the same explanatory variables. In addition residual plots from both models are included where the top panels show the deviance residuals and square roots of the absolute values of standardized deviance residuals against predicted values from the gamma fit whereas the bottom panels gives the corresponding plots for the linear regression.

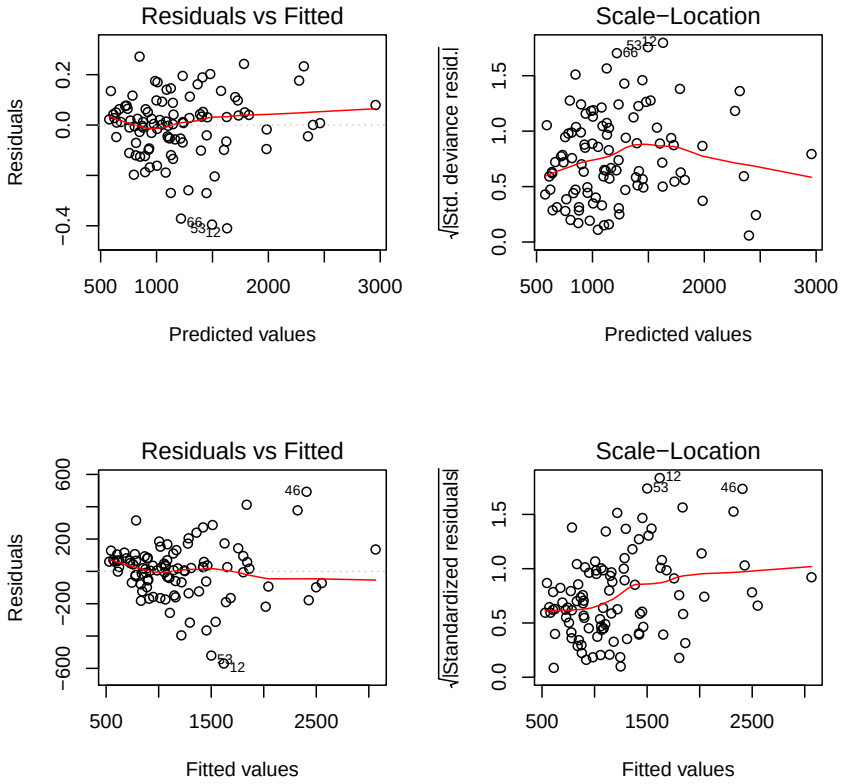
Give a precise statement of the linear models used here.

Discuss differences and similarities between the analyses.

```
Call:
lm(formula = price ~ size + rooms + balcony + rent + x)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  525.90793   70.69386   7.439 4.71e-11 ***
size          20.23111    1.37069  14.760  < 2e-16 ***
rooms          3.76506   37.59353   0.100  0.92044
balcony      129.94298   38.97715   3.334  0.00123 **
rent          -0.13969    0.01704  -8.197 1.23e-12 ***
x           -108.17047   13.46564  -8.033 2.72e-12 ***
---
Residual standard error: 174.9 on 94 degrees of freedom
Multiple R-squared:  0.8948,    Adjusted R-squared:  0.8892
F-statistic: 159.9 on 5 and 94 DF,  p-value: < 2.2e-16
```

(Continued on page 4.)



Problem 3

Let Y_{ij} be count response no. $j = 1, \ldots, d$ in group $i = 1, \ldots, n$. A mixed Poisson model for Y_{ij} with group specific random intercept u_i and one explanatory variable x_{ij} is defined by Y_{ij} being independent and Poisson distributed given u_i with conditional mean

$$\mathrm{E}[Y_{ij}|u_i] = \mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + u_i),$$

thus with a log-link for the mixed model and deterministic $\beta_0 + \beta_1 x_{ij}$.

a) Show that marginally

$$\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij})\mathrm{E}[e^{u_i}]$$

and determine $\mathrm{E}[e^{u_i}]$ when $u_i \sim \mathrm{N}(0, \sigma_u^2)$.

Comment on the relationship between the parameters in the mixed and marginal models.

Hint: The moment generating function of a normal distribution with mean zero and variance σ_u^2 is given by $M(t) = \exp(\sigma_u^2 t^2/2)$.

(Continued on page 5.)

b) Derive an expression for the marginal variance of Y_{ij} under the assumption that $u_i \sim \mathrm{N}(0, \sigma_u^2)$.

c) Assume instead that conditionally on random intercepts $u_i \sim \mathrm{N}(0, \sigma_u^2)$ the responses Y_{ij} are gamma distributed with means $\mathrm{E}[Y_{ij}|u_i] = \exp(\beta_0 + \beta_1 x_{ij} + u_i)$, i.e. still a log-link, and a dispersion term ϕ.

Derive an expression for the marginal mean $\mathrm{E}[Y_{ij}]$ in this situation.

Comment on the relationship between the parameters in the mixed and marginal models also in this situation.

Finally find an expression for the marginal variance of Y_{ij}.

END

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to generalized linear models.
Day of examination:	Friday 14 December 2018.
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

We assume that $V \sim \text{bin}(n, \pi)$, i.e. binomially distributed with n trials and probability of success π, and let $Y = V/n$. Then the probability mass function (pmf) of Y takes the form

$$P(Y = y) = \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \tag{1}$$

for $y = 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$.

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 $c(y, \phi)$.

b) Let μ denote the expected value of Y. Use the expressions for $b(\theta)$ and $a(\phi)$ to show that $\mu = \pi$ and determine $\text{var}(Y)$.

We then assume that $V_1, V_2, \ldots, V_N$ are independent with $V_i \sim \text{bin}(n_i, \pi_i)$, and let $Y_i = V_i/n_i$ for $i = 1, 2, \ldots, N$. We consider a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_N$ with canonical link function $\log\{\pi_i/(1-\pi_i)\} = \eta_i$ and linear predictor $\eta_i = \beta_0 + \beta_1 x_i$. Here $x_1, \ldots, x_N$ are known covariate values.

c) Derive an expression for the log-likelihood function $L(\beta_0, \beta_1)$, and show that the maximum likelihood estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the solutions of the equations

$$\sum_{i=1}^{N} n_i(Y_i - \pi_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{N} n_i x_i (Y_i - \pi_i) = 0,$$

where $\pi_i = e^{\beta_0 + \beta_1 x_i}/(1 + e^{\beta_0 + \beta_1 x_i})$.

(Continued on page 2.)

d) By a general result for maximum likelihood estimation for GLMs, we know that $(\widehat{\beta}_0, \widehat{\beta}_1)^{\mathrm{T}}$ is approximately bivariate normally distributed with mean $(\beta_0, \beta_1)^{\mathrm{T}}$ and a covariance matrix that equals $\boldsymbol{\mathcal{J}}^{-1}$, where $\boldsymbol{\mathcal{J}}$ is the expected information matrix. Find an expression for $\boldsymbol{\mathcal{J}}$.

Problem 2

Titanic was a British passenger liner that sank in the Atlantic Ocean 15 April 1912, after colliding with an iceberg. There were about 1300 passengers and 900 crew aboard, and more than 1500 of them died. In this problem we will use data on 1046 passengers to investigate how the probability of surviving the disaster depends on the age and sex of the passengers and at which class they traveled. (Passengers with no information about age are omitted from the analysis.)

The data file `titanic` that is used in the analysis has one line for each of the 1046 passengers and the following variables in the four columns:

- `Sex`: Sex (1 = male; 2 = female)
- `Cage`: Centered age (age − 30)
- `Class`: Passenger class (1 = first class; 2 = second class; 3 = third class)
- `Survived`: Survived or died (0 = died; 1 = survived)

a) We first fit a model with only main effects, where `Sex` and `Class` are defined to be categorical covariates (factors), while `Cage` is a numeric covariate. The result of this model fit is given below. Describe the model that we have used. Give an interpretation of the estimated intercept and the estimate for (centered) age.

```
Call:
glm(Survived~Sex+Cage+Class, family=binomial, data=titanic)

Coefficients:
               Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)    -0.007567      0.165527     -0.046       0.964
Sex2            2.497845      0.166037     15.044     < 2e-16
Cage           -0.034393      0.006331     -5.433    5.56e-08
Class2         -1.280571      0.225538     -5.678    1.36e-08
Class3         -2.289661      0.225802    -10.140     < 2e-16

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
```

b) On the next page is given the result for a model with interaction between sex and passenger class. Explain why this model is to be preferred to the one in question a. Describe which effects sex and passenger class have for the probability of surviving the disaster.

(Continued on page 3.)

```
Call:
glm(Survived~Sex+Cage+Class+Sex:Class, family=binomial, data=titanic)

Coefficients:
                  Estimate   Std. Error    z value     Pr(>|z|)
(Intercept)       -0.234083     0.186230     -1.257       0.209
Sex2               3.886388     0.492375      7.893    2.95e-15
Cage              -0.038401     0.006743     -5.695    1.23e-08
Class2            -1.600280     0.301987     -5.299    1.16e-07
Class3            -1.576159     0.252514     -6.242    4.32e-10
Sex2:Class2        0.070407     0.630978      0.112       0.911
Sex2:Class3       -2.488805     0.540041     -4.609    4.05e-06

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  931.99  on 1039  degrees of freedom
```

c) Finally we fit models with more interactions. A summary of the fits of these models is given in the analysis of deviance table below. Some of the numbers in the table have been replaced by question marks. Fill in the correct numbers for the question marks, and explain how you arrive at these numbers. Which of the four models would you prefer? (Give the reasons for your answer.)

```
> anova(fit1,fit2,fit3,fit4,test="LRT")

Analysis of Deviance Table

Model 1: Survived~Sex+Cage+Class+Sex:Class
Model 2: Survived~Sex+Cage+Class+Sex:Class+Cage:Class
Model 3: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage
Model 4: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage+Sex:Cage:Class
     Resid. Df    Resid. Dev    Df    Deviance    Pr(>Chi)
1         1039        931.99
2         1037        922.17     ?        ?        0.00739
3           ?         917.84     1      4.3308     0.03743
4         1034           ?       2      1.8661     0.39335
```

Problem 3

Let $Y_1, \ldots, Y_n$ be independent and normally distributed with common variance σ^2 and

$$\mu_i = E(Y_i) = \beta_0 + \beta_1(x_i - \bar{x}); \qquad i = 1, 2, \ldots, n; \tag{3}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. We introduce the vectors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$, and write (3) on vector/matrix form

$$\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Here $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ and $\boldsymbol{X} = [\mathbf{1}_n\,,\, \boldsymbol{x} - \bar{x}\mathbf{1}_n]$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{1}_n$ is a n-dimensional vector of 1's.

(Continued on page 4.)

a) Show that the projection matrix onto the model space $C(\mathbf{X})$ takes the form

$$\boldsymbol{P}_1 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}},$$

where $M = \sum_{i=1}^n (x_i - \bar{x})^2$.

b) Use the result in question a to show that the vector of fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_1\boldsymbol{Y}$ may be written

$$\widehat{\boldsymbol{\mu}} = \overline{Y}\mathbf{1}_n + \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n),$$

where $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\widehat{\beta}_1 = M^{-1}\sum_{i=1}^n Y_i(x_i - \bar{x})$.

If $\beta_1 = 0$ we have the null model with $\mu_i = E(Y_i) = \beta_0$ for $i = 1, \ldots, n$. The projection matrix for the null model is known to be $\boldsymbol{P}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$. We then have the orthogonal decomposition

$$\boldsymbol{Y} = \boldsymbol{P}_0\boldsymbol{Y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}$$

with corresponding sum of squares decomposition

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}. \tag{4}$$

c) Show that

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} = M\widehat{\beta}_1^2,$$

and

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} = \sum_{i=1}^n \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2.$$

d) Use Cochran's theorem to show that

$$M\widehat{\beta}_1^2/\sigma^2 \tag{5}$$

and

$$\sum_{i=1}^n \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2/\sigma^2 \tag{6}$$

are independent and (non-central) chi-squared distributed. Determine the degrees of freedom for the statistics (5) and (6), and show that the non-centrality parameter of (6) is 0. (One may show that the non-centrality parameter of (5) equals $M\beta_1^2/\sigma^2$, but you should not show this.)

e) Derive an F-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_A : \beta_1 \neq 0$, and determine the distribution of the test statistic under H_0 and under H_A.

END

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

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Wednesday November 30th 2016
Examination hours:	14.30 – 18.30

This problem set consists of 4 pages.

Appendices:	Tables for normal-, t-, χ^2-distributions
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

In this problem you shall consider models where the response is considered as gamma distributed, i.e. $G(\mu, \nu)$. The density function is

$$f(y; \mu, \nu) = \frac{y^{-1}}{\Gamma(\nu)}(\frac{y\nu}{\mu})^{\nu} \exp(-y\nu/\mu),\ y > 0.$$

a) Express this as an exponential distribution where the density function has the form

$$c(y, \phi) \exp(\frac{\theta y - a(\theta)}{\phi}).$$

Identify $\theta, \phi, a(\theta)$ and $c(y, \phi)$.

b) Explain what the canonical link function is in this case, and discuss its use.

Problem 2

In the year 2014 147 persons were killed in road accidents in Norway. The figures classified according to gender and eight age groups can be found at the end of the problem set together with the size of the population in each group.

The output from fitting a model assuming that the responses were Poisson distributed

```
mod1<-glm(killed~factor(gender) + factor(age) +offset(log(population/100000),
                family=poisson,data=accidents)
```

is displayed below. The factors `gender` with male as base level and `age` group with 0-17 as base level, are used as covariates. The link function is the canonical link. Remark also that in the command the population size is divided by 100 000, so rates must be interpreted per 100 000 individuals.

(Continued on page 2.)

```
Call:
glm(formula = killed ~ factor(gender) + factor(age)
+ offset(log(population/1e+05)),family = poisson, data = accidents)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09971  -0.40719  -0.00551   0.57037   1.05081

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.1506     0.3367   0.447 0.654634
factor(gender)2  -1.0212     0.1858  -5.495 3.91e-08 ***
factor(age)2      1.5565     0.4082   3.813 0.000137 ***
factor(age)3      1.0102     0.4216   2.396 0.016584 *
factor(age)4      0.8869     0.4272   2.076 0.037918 *
factor(age)5      1.5366     0.3867   3.973 7.09e-05 ***
factor(age)6      1.3329     0.4082   3.265 0.001095 **
factor(age)7      1.9813     0.3868   5.123 3.01e-07 ***
factor(age)8      2.1126     0.3988   5.297 1.18e-07 ***
---


(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 92.6417  on 15  degrees of freedom
Residual deviance:  9.9502  on  7  degrees of freedom
```

a) Discuss why a Poisson model is reasonable in this case. Using the information from the output, how would you judge the model fit?

b) Why is it sensible to include an `offset` in the linear predictor in this case?

c) What is the interpretation of the estimate of the intercept, β_0? What is the p-value of a test that the coefficient of the gender effect is equal to -1?

d) What is the fitted value and residual for women in the age group 45-54 years?

e) The total number of women who were killed was 40. The sum of the fitted values for women of all age group is also 40. Explain why.

Problem 3

The data in this problem is based on four measurements of a particular bone for 5 randomly selected boys. The measurements were taken at 8, 8.5, 9 and 9.5 years.

The variables included are

- bone: length of the bone in millimeters

(Continued on page 3.)

- redage: age -8.75, i.e. age centered

Below is an excerpt from the output from fitting a linear mixed model (LMM) with the procedure `lme` in R where the length of the bone, `bone` is the response,

$$\texttt{bone}_{ij} = \beta_0 + \beta_1 \texttt{redage}_{ij} + b_{i,1} + \texttt{redage}_{ij} b_{i,2} + \varepsilon_{ij}, i = 1, \ldots, 5,\ j = 1, 2, 3, 4$$

where $\mathbf{b}_i = (b_{i,1}, b_{i,2})^T,\ i = 1, \ldots, 5$ represent the random effects.

```
Linear mixed-effects model fit by REML
 Data: bonedat
       AIC      BIC    logLik
  44.97205 50.31428 -16.48603

Random effects:
 Formula: ~1 + redage | boy
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.8172867 (Intr)
redage      0.7323611 0.586
Residual    0.2939400

Fixed effects: bone ~ 1 + redage
             Value Std.Error
(Intercept) 52.690 0.3713644
redage       1.424 0.3479866
 Correlation:
       (Intr)
redage 0.543

Number of Observations: 20
Number of Groups: 5
```

a) Formulate the model on matrix form and explain the meaning and interpretation of the different parts. State the usual model assumptions carefully.

b) Use the values in the R-output to describe the distribution of the response $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^T,\ i = 1, \ldots, 5$.

c) Find the conditional expectation of a random effect $\mathbf{b}_i,\ i = 1, \ldots, 5$ given the observations, i.e. $E[\mathbf{b}_i | \mathbf{y}_1, \ldots, \mathbf{y}_5]$. Describe how the random effects, $\mathbf{b}_i,\ i = 1, \ldots, 5$, can be predicted/estimated.

In part b) and c) it is not necessary to do any numerical calculations.

(Continued on page 4.)

Observed number of killed in road traffic in 2014 and size of Norwegian population according to gender, 1= male, 2= female, and age 1=0-17 , 2=18-24, 3=25-34, 4=35-44, 5=45-54, 6=55-64, 7=65-74, 8=75+

	population	killed	gender	age
1	576584	5	1	1
2	243510	11	1	2
3	349669	10	1	3
4	370541	11	1	4
5	359178	24	1	5
6	301981	12	1	6
7	224471	19	1	7
8	141500	15	1	8
9	548577	4	2	1
10	230407	7	2	2
11	333730	5	2	3
12	348593	3	2	4
13	338505	2	2	5
14	295223	6	2	6
15	232997	7	2	7
16	213610	6	2	8

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Monday November 30th 2015
Examination hours:	14.30 – 18.30

This problem set consists of 4 pages.

Appendices:	Tables for normal-, t-, χ^2-distributions
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

In this problem you shall consider models where the response is considered as binomially distributed $Bin(n, \pi)$. Let $\mu = n\pi$.

a) Express this as a generalized linear model where the frequency distribution has the form

$$c(y, \phi) \exp(\frac{\theta y - a(\theta)}{\phi}).$$

Explain what is meant by a link function.

b) Assume $y_i, n_i, x_{i,1}, \ldots, x_{i,p+1}$, $i = 1, \ldots, n$ are n set of observations where y_i are are the responses and $x_{i,1}, \ldots, x_{i,p+1}$ are the covariates. The responses are assumed to be independent $Bin(n_i, \pi_i)$ distributed. Let $\hat{\mu}_i = n_i \hat{\pi}_i$ be the fitted values. What is the deviance from fitting this model? How is it expressed in this case? What is the main use of the deviance?

Problem 2

In this problem you shall consider data of survivals from a study of treatment for breast cancer. The response is the numbers that survived for three years. The covariates were the four factors

- app: appearance of tumor, two levels 1=malignant, 2=benign
- infl: inflammatory reaction, two levels 1= minimal, 2=moderate or severe
- age: age of patients, three levels 1= under 50, 2= 50-69, 3=70 or older
- country: hospital of treatment, three levels, 1= Japan, 2= US, 3= UK

(Continued on page 2.)

The number of survivors is modeled as a binomially distributed variable using a canonical logit link. Level 1 is used as base level or reference category for all factors.

a) The output from fitting the model where only appearance and country are used as covariates, i.e. a model with predictor of the form

$$\eta = \beta_0 + \beta_1 \text{fapp} + \beta_2 \text{fcountry2} + \beta_3 \text{fcountry3}$$

is displayed below. What is the interpretation of the estimate of the coefficient of appearance, `fapp` (`f` means factor)? Explain also how the coefficient can be expressed in terms of an odds ratio.

```
Call:
glm(formula = cbind(surv, nsurv) ~ fapp + fcountry, family = binomial,
    data = brc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8033  -0.7267   0.2157   0.7579   1.8742

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0811     0.1656   6.529 6.63e-11 ***
fapp2         0.5140     0.1659   3.098 0.001949 **
fcountry2    -0.6616     0.1993  -3.319 0.000902 ***
fcountry3    -0.4946     0.2071  -2.389 0.016917 *
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.637  on 35  degrees of freedom
Residual deviance: 36.662  on 32  degrees of freedom
```

b) The output below is an analysis of deviance table for comparing various model specifications. Fill out the positions indicated by a question mark.

(Continued on page 3.)

```
Analysis of Deviance Table

Model 1: cbind(surv, nsurv) ~ fapp + fage + fcountry
Model 2: cbind(surv, nsurv) ~ fapp + fage + finfl + fcountry
Model 3: cbind(surv, nsurv) ~ fapp + finfl + fage * fcountry
Model 4: cbind(surv, nsurv) ~ fapp * finfl + fage * fcountry
Model 5: cbind(surv, nsurv) ~ fapp * finfl + fapp * fage + fage * fcountry
Model 6: cbind(surv, nsurv) ~ fapp * finfl * fage * fcountry
  Resid. Df Resid. Dev Df Deviance
1        30     33.198
2         ?     33.197  1   0.0009
3        25     25.718  ?   7.4790
4        24     25.511  1        ?
5        22     22.059  2   3.4519
6         0      0.000  ?  22.0587
```

In the remaining parts of this problem we return to the model in part a) and consider the hypothesis

$$H_0 : \beta_2 + \beta_3 = -1 \text{ versus } H_a : \beta_2 + \beta_3 \neq -1.$$

c) The estimated covariance matrix between the estimators of the coefficients β_2 and β_3 is $\begin{pmatrix} 0.040 & 0.021 \\ 0.021 & 0.043 \end{pmatrix}$. Use a Wald test to test the null hypothesis above.

d) Explain how the null hypothesis can be tested with a likelihood ratio test by fitting two suitable models. No numerical calculations are necessary, but it must be specified how the pedictors should be defined.

Problem 3

The data used in this problem is for expenses in the the social security system Medicare in US. Average expenses per hospitalization, denoted as `ccpd`, were in six years recorded for 54 regions: the fifty states, Puerto Rico, Virgin Islands, District of Columbia and an unspecified other. Thus there are $6 \times 54 = 324$ observations. The expenses are treated as response. The covariates are `j = YEAR` which can take values $1, \ldots, 6$ and a factor indicating the average length of stay at hospital, `AVETD` in each region and year. This factor has tree levels, 1= six days or less, 2= 7-9 days, 3= 10 days or more. Six days or less is the base level and the others are denoted as AVETD_2 and AVETD_3.

Below the output from fitting the linear mixed effects model

$$y_{ij} = \beta_0 + \beta_1 \times j + \beta_2 \text{AVETD}_{2ij} + \beta_3 \text{AVETD}_{3ij} + b_{1i} + j \times b_{2i} + \varepsilon_{ij}, \quad j = 1, \ldots, 6 \, , i = 1, \ldots, 54$$

is displayed

(Continued on page 4.)

```
Linear mixed-effects model fit by REML
 Data: medicare
      AIC      BIC   logLik
  5200.98 5231.127 -2592.49

Random effects:
 Formula: ~1 + YEAR | fstate
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2410.7972 (Intr)
YEAR         262.7191 0.418
Residual     429.6119

Fixed effects: ccpd ~ YEAR + factor(AVETD)
                   Value Std.Error  DF   t-value p-value
(Intercept)    7419.853  386.0518 267 19.219839  0.0000
YEAR            706.045   39.5543 267 17.849996  0.0000
factor(AVETD)2  567.721  183.9157 267  3.086857  0.0022
factor(AVETD)3 1008.339  244.2480 267  4.128342  0.0000
 Correlation:
               (Intr) YEAR   f(AVETD)2
YEAR            0.170
factor(AVETD)2 -0.488  0.168
factor(AVETD)3 -0.468  0.239  0.781

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.19288486 -0.60341726  0.01798739  0.61830554  3.51342658

Number of Observations: 324
Number of Groups: 54
```

a) Formulate the model in matrix form and explain what the usual assumptions are.

b) Compute a 95% confidence interval for the fixed effect coefficient for `YEAR`.

c) Explain how a test for simplifying the model by removing the random effect $b2$ can be performed.

d) What is the expectation and covariance matrix in the marginal model, i.e. of the response $(y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6})'$?

e) Explain how the null hypothesis $H_0 : \beta_3 = 2 \times \beta_2$ versus the alternative hypothesis $H_a : \beta_3 \neq 2 \times \beta_2$ can be tested? In this part no numerical calculations are expected.

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Monday December 1th 2014
Examination hours:	14.30 – 18.30
This problem set consists of 6 pages.	
Appendices:	None
Permitted aids:	Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Each subtask indexed by letters (1a, 1b etc.) counts equally. Each question numbered with Roman numerals (i), ii) and iii)) counts equally within each subtask.

Problem 1

1a

A distribution belongs to the exponential family if its probability mass function or probability density can be written in the form

$$f(y;\theta,\phi) = c(y,\phi)\exp[(\theta y - a(\theta))/\phi],$$

where $a(\cdot)$ and $c(\cdot,\cdot)$ are functions.

i) Show that if Y is a stochastic variable with a distribution belonging to the exponential family, then $\mathrm{E}(Y) = a'(\theta)$ and $\mathrm{Var}(Y) = \phi a''(\theta)$, where a' and a'' denote the first and second derivatives of a. [Hint: Start with calculating the first derivative of $f(y;\theta,\phi)$ with respect to θ.]

1b

The probability mass function for a Poisson distributed variable Y is

$$f(Y = y;\lambda) = (\lambda^y/y!)\exp(-\lambda).$$

i) Show that the Poisson distribution belongs to the exponential family.

ii) Show that $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

(Continued on page 2.)

1c

Consider a regression problem with a Poisson distributed response variable Y, with logarithmic link function and with two explanatory variables x_1 and x_2 such that

$$Y \sim \text{Po}(\mu),$$
$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

i) Give an interpretation of the parameter β_1 or some transformation of it.

ii) Assume then that $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 3$ and predict the response Y for $x_1 = 1$ and $x_2 = 1$ and then for $x_1 = 2$ and $x_2 = 1$.

1d

Consider now a specific data set with 100 observations of a count variable Y and two explanatory variables x_1 and x_2. The model in the previous exercise has been fitted to these data. Below you see some R output with information about the fitted model.

```
summary(glmobj)

Call:
glm(formula = y ~ x1 + x2, family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-15.0942   -0.7773   -0.3345    0.5244   10.9833

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.863643   0.031640    27.3   <2e-16 ***
x1          2.132696   0.007939   268.6   <2e-16 ***
x2          2.970372   0.012227   242.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 264872.7  on 99  degrees of freedom
Residual deviance:   1063.8  on 97  degrees of freedom
AIC: 1372.6
Number of Fisher Scoring iterations: 4

> phihat<-sum(residuals(glmobj,type="pearson")^2)/(100-3)
> phihat
[1] 11.32317
```

i) Explain what over-dispersion means in Poisson regression.

ii) Explain why the results above show that the current count data are over-dispersed.

iii) Discuss shortly two different possibilities for performing a more correct analysis than that given above.

(Continued on page 3.)

Problem 2

2a

i) Give an interpretation of a regression coefficient β, or a transformation of it, in binary regression with logit link function.

ii) Give then a simpler interpretation of β which holds approximately for small probabilities.

2b

Consider a situation with 50 observations of a binary response variable Y and two continuous explanatory variables x_1 and x_2, where we fit models with different link functions and different explanatory variables included. Below you see the R code for fitting ten different models and the corresponding values of Akaike's Information Criterion (AIC).

```
> m0<-glm(y~1,family=binomial(link=log))
> m1.logit<-glm(y~x1,family=binomial(link=logit))
> m2.logit<-glm(y~x2,family=binomial(link=logit))
> m12.logit<-glm(y~x1+x2,family=binomial(link=logit))
> m1.probit<-glm(y~x1,family=binomial(link=probit))
> m2.probit<-glm(y~x2,family=binomial(link=probit))
> m12.probit<-glm(y~x1+x2,family=binomial(link=probit))
> m1.cloglog<-glm(y~x1,family=binomial(link=cloglog))
> m2.cloglog<-glm(y~x2,family=binomial(link=cloglog))
> m12.cloglog<-glm(y~x1+x2,family=binomial(link=cloglog))

> AIC(m0,
    m1.logit,m2.logit,m12.logit,
    m1.probit,m2.probit,m12.probit,
    m1.cloglog,m2.cloglog,m12.cloglog)

            df      AIC
m0           1 68.40641
m1.logit     2 55.38840
m2.logit     2 70.26156
m12.logit    3 57.11896
m1.probit    2 55.37494
m2.probit    2 70.26544
m12.probit   3 57.15451
m1.cloglog   2 55.62994
m2.cloglog   2 70.28034
m12.cloglog  3 57.56906
```

i) Define AIC.

ii) Which one of the models above would you choose based on the given results? Why?

2c

Assume that the data above are used to calibrate a test for diagnosing a disease, such that we predict that a patient has a disease ($Y = 1$) if the

(Continued on page 4.)

predicted probablity is larger than a threshold value γ.

i) Define the two terms sensitivity and specificity.

ii) Describe what a ROC (Receiver Operating Characteristics) curve is, and draw a plot with one curve for a model with good classification performance and another model which is no better than random classification.

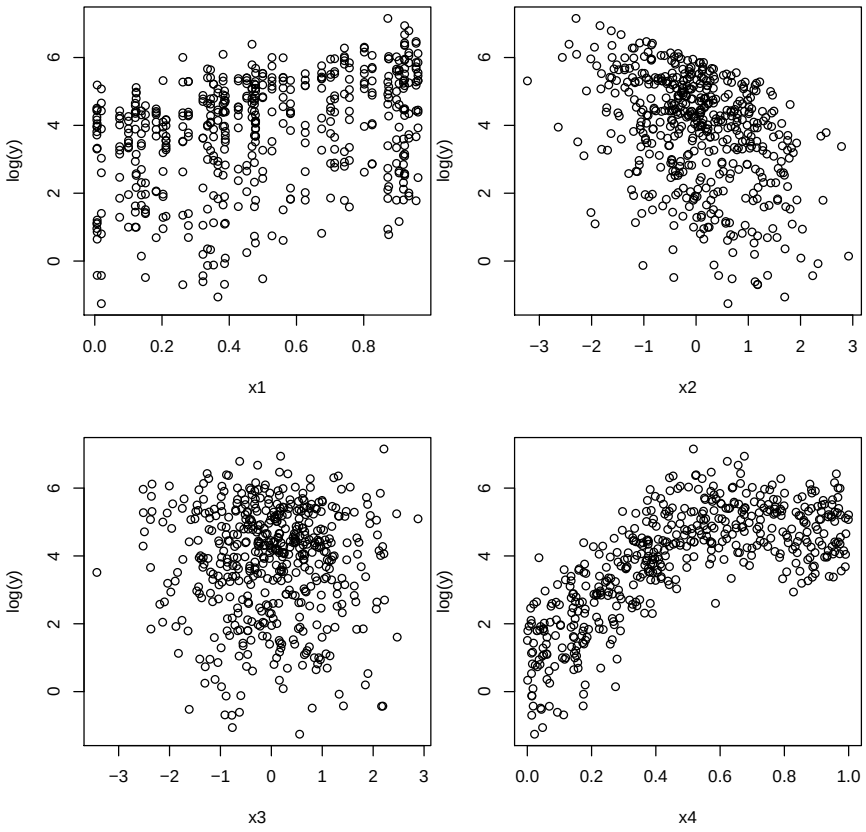
Problem 3

Consider a regression problem where

- The response Y is a continuous positive variable
- Y and corresponding explanatory variables are observed for 50 different groups with 10 observations within each group
- The continuous explanatory variable x_1 is group specific and has the same value within each group
- The three continuous explanatory variables x_2, x_3 and x_4 may have different values both between groups and between observations within the same group
- $\text{Var}(x_1) = 0.087$, $\text{Var}(x_2) = 0.98$, $\text{Var}(x_3) = 1.06$ and $\text{Var}(x_4) = 0.086$
- Groups are indexed by $i, i = 1, \ldots 50$ and observations within each group by $j, j = 1, \ldots 10$

(Continued on page 5.)

Below are scatter plots of the logarithm of the response vs. each of the explanatory variables.



We assume that the groups are a random subset of a population of groups. The following model has been fitted to these data

$$\begin{aligned} Y_{ij} &\sim \text{Gamma}(\mu_{ij}, \phi), \\ \log(\mu_{ij}) &= \beta_0 + b_i + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}, \\ b_i &\sim N(0, \sigma_b^2), \end{aligned}$$

using the R code

```
> require(lme4)
> glmmobj<-glmer(y~x1+x2+x3+x4 + (1|g), family=Gamma(link=log))
```

Below you see a summary of the fitted object:

```
> summary(glmmobj)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: Gamma  ( log )
Formula: y ~ x1 + x2 + x3 + x4 + (1 | g)

     AIC      BIC   logLik deviance df.resid
  4834.0   4863.5  -2410.0   4820.0      493
```

(Continued on page 6.)

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5533 -0.8675 -0.2161  0.7478  3.6526

Random effects:
 Groups   Name        Variance Std.Dev.
 g        (Intercept) 0.006691 0.0818
 Residual             0.353262 0.5944
Number of obs: 500, groups:  g, 50

Fixed effects:
            Estimate Std. Error t value Pr(>|z|)
(Intercept)  1.04721    0.08737   11.99   <2e-16 ***
x1           2.06204    0.10670   19.33   <2e-16 ***
x2          -0.64944    0.02861  -22.70   <2e-16 ***
x3           0.02241    0.02685    0.83    0.404
x4           4.14268    0.12616   32.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
   (Intr) x1     x2     x3
x1 -0.625
x2 -0.038  0.049
x3 -0.019  0.013 -0.049
x4 -0.710  0.032  0.008  0.004
```

3a

i) Discuss whether the random effect term b_i is an important part of the model compared to other parts of the model.

3b

i) Use the information you have to suggest simplifications or improvements of the model.

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100/STK4100 — Introduction to generalized linear models

Day of examination: Friday December 6th 2013

Examination hours: 14.30 – 18.30

This problem set consists of 3 pages.

Appendices: Table over normal distribution and table over χ^2-distribution

Permitted aids: Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Problem 1

a) Densities and frequency functions of the form

$$f(y;\theta,\phi) = c(y;\phi)\exp(\frac{\theta y - a(\theta)}{\phi})$$

to describe the distribution of the response variable Y are used to define a generalized linear model (GLM). Describe the other parts of a GLM, and how they are related to $f(y;\theta,\phi)$.

b) Find a general expression for the expectation and the variance of the response variable Y.

c) What is meant with a saturated model? Explain how the deviance in a GLM is defined and explain how it can be used to compare two models. How is this procedure related to likelihood ratio tests? How are the deviance residuals defined?

d) Now consider the Poisson distribution and find expressions for the deviance and deviance residuals in this case.

Table 1: Observed numbers.

Number of children	0	1	2	3	4	5	6	Total
7 and 10 years educ.	128	161	194	61	12	6	2	564
13, 16 and 19 years educ.	44	26	37	5	2	1	0	115
All women	172	187	231	66	14	7	2	679

In the table above the number of children of 679 German women is recorded. The total number of children is 950. To investigate if there is any relation between years of education and number of children the material was divided into two parts: one with women with 7 and 10 years of education and

(Continued on page 2.)

another with those having 13, 16 or 19 years of education. Then two models were fitted using a GLM with Poisson distributed response and the canonical `log` link : one without any covariates, so the predictor was the same for all women, and another where a factor with two levels, that indicated to which groups the woman belonged, was the covariate. The number of children for each woman was considered as the response. The fitted values were 1.399 for the first model. For the second model the fitted values were 1.457 for those with 7 or 10 years education and 1.113 for the other.

e) Explain how the deviance can be used to test if there is any difference between the two groups with respect to the number of children. Carry out the test. Use a 5% level.

f) If $y_i,\ i = 1, \ldots, 679$ are the observed values and $\hat{\mu}_i,\ i = 1, \ldots, 679$ are the fitted values, explain why $\sum_{i=1}^{679}(y_i - \hat{\mu}_i) = 0$ for both models. Is this result also true for more general models?

Problem 2

The data in this problem is part of a longitudinal study of income in the US, the Panel Study of Income Dynamics, begun in 1968. The subset consists of 42 heads of household who were aged 25-39 in 1968. The variables included are

- annual nominal income, which is the response variable
- age, age in 1968
- cyear, coded as -10 in 1968,0 in 1978 and 10 in 1988
- educ, years of education in 1968
- sex, M=male, F=female

Below is an excerpt from the output from fitting a linear mixed model (LMM) with the procedure `lme` in R where log income, `lincm` is the response,

$$\texttt{lincm}_{ij} = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{cyear}_{ij} + \beta_3 \texttt{educ}_i + \beta_4 \texttt{sex}_i + b_i + \varepsilon_{ij}, i = 1, \ldots, 42,\ j = 1968, 1978, 1988$$

where b_i represent the random effects.

```
Linear mixed-effects model fit by REML
 Data: psid2
       AIC      BIC    logLik
  320.0178 339.5883 -153.0089

Random effects:
 Formula: ~1 | fid
        (Intercept)  Residual
StdDev:   0.0419192 0.7505293
```

(Continued on page 3.)

```
Fixed effects: lincm ~ age + cyear + educ + factor(sex)
                 Value Std.Error  t-value
(Intercept)   7.386823 0.6317104  11.693370
age          -0.020930 0.0152069  -1.376316
cyear         0.084163 0.0081889  10.277583
educ          0.116343 0.0275823   4.218021
factor(sex)M  1.311661 0.1422471   9.221007

Correlation:
             (Intr) age    cyear  educ
age          -0.831
cyear         0.000  0.000
educ         -0.685  0.201  0.000
factor(sex)M  0.003 -0.217  0.000  0.041

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.4411400 -0.4003431  0.1070887  0.5602338  1.6037724

Number of Observations: 126
Number of Groups: 42
```

a) Formulate the model on matrix form and explain the meaning of the different parts. State the model assumptions carefully.

b) Determine an approximate 95% interval for the coefficient of `cyear`. Do you think the nominal income has been constant in the period covered by the survey?

c) If one is interested in the simultaneous significance of two fixed effects, `age` and `educ` say, describe how that can be tested in this kind, LMM, of models.

d) Describe how the random effects b_i can be predicted/estimated?

e) Use the values in the R-output to calculate the estimated covariance matrix for the response $(Y_{i1}, Y_{i2}, Y_{i3})^T$.

In part c) and d) it is not necessary to do any numerical calculations.

END

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100/4100 — Innføring i generaliserte lineære modeller
Eksamensdag:	Torsdag 6. desember 2012.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over χ^2 og t fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

De ulike delpunktene kan stort sett løses uavhengige av hverandre. Hvis du står fast på et punkt, gå derfor heller videre til neste punkt.

Oppgave 1

En stokastisk variabel Y sies å ha fordeling i den eksponensielle fordelingsklasse dersom tettheten (eller punkt sannsynligheten) til Y kan skrives på formen

$$f(y;\theta,\phi) = c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi}).$$

For videre utregninger får du oppgitt at hvis

$$M_Y(t) = \mathrm{E}[\exp(Yt)] = \int \exp(yt) f(y) dy$$

eksisterer for alle t i et omegn om 0, så er

$$\mathrm{E}[Y^r] = M_Y^{(r)}(0)$$

der $M_Y^{(r)}(\cdot)$ er den r-te deriverte av $M_Y(t)$ mhp t.

(a) Regn ut forventning og varians i den eksponensielle fordelingsklasse.

Vi vil i resten av denne oppgaven se på den inverse Gaussiske fordeling, gitt ved

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma}} \exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \quad y > 0$$

(Fortsettes på side 2.)

(b) Vis at denne fordelingen tilhører den eksponensielle familie og vis at $\theta = -1/(2\mu^2)$ og $a(\theta) = -\sqrt{-2\theta}$. Identifiser også ϕ og $c(y;\phi)$.

(c) Finn forventning og varians i den inverse Gaussiske fordeling. Bruk dette til å diskutere i hvilke situasjoner en slik fordeling kan være nyttig å bruke.

Hva slags begrensninger ligger det på parametrene som er involvert?

(d) Anta nå $Y_1, ..., Y_n$ er uavhengige variable fra en generalisert lineær modell (GLM) med invers Gaussisk fordeling som respons fordeling. Forklar hva dette betyr.

Forklar generelt hva devians betyr og diskuter hva devians kan brukes til i GLM-sammenheng.

(e) Forklar hva vi mener med kanonisk link og hvilke fordeler det har å bruke denne.

Hva blir den kanoniske link for den inverse Gaussiske fordeling?

Oppgave 2

Vi skal i denne oppgaven se på modeller med følgende struktur:

$$\begin{aligned} Y_{ij} =& \beta_0 + b_{0,i} + (\beta_1 + b_{1,i})x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \\ \boldsymbol{b}_i =& (b_{0,i}, b_{1,i})^T \sim N(\mathbf{0}, \boldsymbol{D}) \end{aligned} \tag{*}$$

der $i \in \{1, ..., N\}$ er en gruppe-indeks mens $j \in \{1, ..., n_i\}$ er en indeks for repeterte målinger innen gruppe. Vi antar her at alle b- og ε-variable er uavhengige av hverandre.

(a) Hva kalles denne modellen? (Bruk gjerne det engelske navnet.)

Diskutér nytten av slike modeller.

(b) Hva blir den *marginale* modellen for $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$?

Hvilke fordeler har det at vi har et eksplisitt uttrykk for den marginale fordelingen til $\mathbf{Y}_i$ når det gjelder estimering?

(c) Forklar hovedprinsippene ved REML estimering.

Diskutér fordeler og ulemper med maksimum likelihood (ML) estimering sammenliknet med REML estimering. Spesifiser spesielt i hvilke tilfeller en vil bruke de ulike metodene.

Davidian og Giltinan (1995) beskriver et datasett for å sammenlikne vekstmønstre for to typer av soyabønner. Datasettet består av 412 observasjoner fordelt på 48 jordstykker med 8-10 observasjoner innen hvert jordstykke. I tillegg til vekt (`weight`) og type soyabønne (`Variety`, to typer)

(Fortsettes på side 3.)

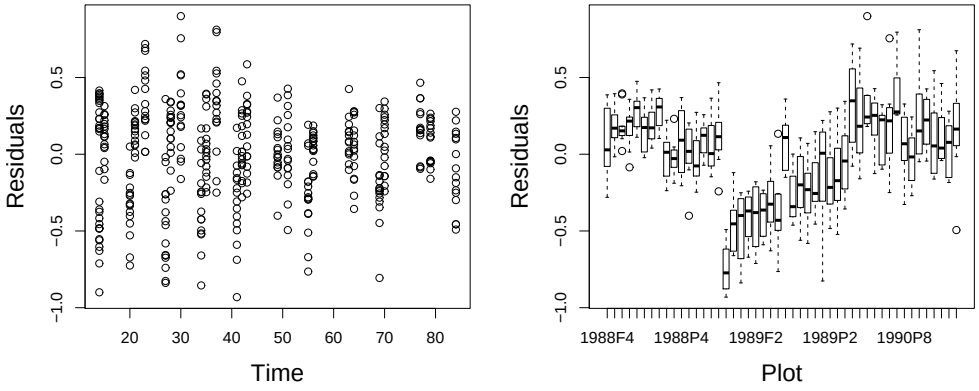
er også tidspunkt for innsamling av observasjon (dager etter planting, `Time`) angitt. Variabelen `Plot` angir jordstykke. I tillegg innfører vi variabelen `Time2` som er `Time` kvadrert.

(d) Vi vil først se på en enkel modell der vekt på log-skala er brukt som responsvariabel mens `Variety`, `Time` og `Time2` er inkludert som forklaringsvariable i tillegg til interaksjon mellom `Variety` og `Time`. Vi skriver denne modellen generelt som

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \varepsilon_{ij} \tag{M0}$$

der i angir jordstykke mens j angir replikasjon innen jordstykke.

Figuren nedenfor viser boksplot av residualer gruppert etter jordstykker og plot av residualer mot `Time`. Kommentér plottene og argumenter hvorfor en modell tilsvarende $(*)$ kan være nyttig i dette tilfellet.



(e) Modellen i foregående deloppgave er så utvidet til to alternative modeller

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij} \tag{M1}$$

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_{0,i} + b_{1,i}\text{Time}_{ij} + \varepsilon_{ij} \tag{M2}$$

der $b_i \sim N(0, d^2)$ i modell M1 og $\boldsymbol{b}_i = (b_{0,i}, b_{1,i})^T \sim N(\boldsymbol{0}, \boldsymbol{D})$ i modell M2. Log-likelihood verdiene (innsatt REML estimater) for de tre modellene er gitt nedenfor:

Modell	M0	M1	M2
Loglik	-130.92	-13.42	6.16

Basert på dette, begrunn hvorfor modell M2 er å foretrekke.

(f) Nedenfor er resultatet av en tilpasning av modell M2 gjort med ML estimering. Diskutér om det er behov for å forenkle modellen.

```
Linear mixed-effects model fit by maximum likelihood

Random effects:
```

(Fortsettes på side 4.)

```
 Formula: ~1 + Time | Plot
            StdDev      Corr
(Intercept) 0.373119650 (Intr)
Time        0.002970683 -0.999
Residual    0.190066092

Fixed effects: log(weight) ~ Variety + Time + Time2 + Variety:Time
                   Value  Std.Error  DF   t-value p-value
(Intercept)   -5.202444 0.09214869 361 -56.45707  0.0000
VarietyP       0.478989 0.11677518  46   4.10181  0.0002
Time           0.204265 0.00236086 361  86.52125  0.0000
Time2         -0.001317 0.00002330 361 -56.52636  0.0000
VarietyP:Time -0.003079 0.00124599 361  -2.47102  0.0139
```

SLUTT

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in	STK3100/4100 — Introduction to generalized l
Day of examination:	Thursday 6. desember 2012.
Examination hours:	14.30 – 18.30.
This problem set consists of 4 pages.	
Appendices:	Tabell over normal, χ^2 og t fordeling
Permitted aids: STK1100/STK1110 and STK2120	Accepted calculator. Formulae notes for

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

The different sub-points can mainly be solved independently. If you get stuck at one point, go further to the next point.

Problem 1

A stochastic variable Y follows a distribution in the exponential distribution family if the density (or the probability) for Y can be written as

$$f(y;\theta,\phi) = c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi}).$$

For further calculations you can use that if

$$M_Y(t) = \mathrm{E}[\exp(Yt)] = \int \exp(yt) f(y) dy$$

exist for all t in a neighborhood of 0, then

$$\mathrm{E}[Y^r] = M_Y^{(r)}(0)$$

where $M_Y^{(r)}(\cdot)$ is the r-th derivative of $M_Y(t)$ wrt t.

(a) Calculate the expectation and variance in the exponential distribution class.

We will in the rest of this exercise look at the inverse Gaussian distribution, given by

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma}} \exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \quad y > 0$$

(Continued on page 2.)

(b) Show that this distribution belongs to the exponential family and show that $\theta = -1/(2\mu^2)$ and $a(\theta) = -\sqrt{-2\theta}$. Also identify ϕ and $c(y;\phi)$.

(c) Find the expectation and the variance in the inverse Gaussian distribution. Use this to discuss what situations such a distribution can be useful to apply.

What kind of constraints are there in the parameters involved?

(d) Assume now $Y_1, ..., Y_n$ are independent variables from a generalized linear model (GLM) with the inverse Gaussian distribution as response distribution. Explain what this means.

Explain in general what deviance means and discuss what deviance can be used to in a GLM setting.

(e) Explain what we mean by canonical link and what kind of advantages there is in using such link functions.

What is the canonical link for the inverse Gaussian distribution?

Problem 2

We will in this exercise look at the models of the following type:

$$\begin{aligned} Y_{ij} =& \beta_0 + b_{0,i} + (\beta_1 + b_{1,i})x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \qquad (*) \\ \boldsymbol{b}_i =& (b_{0,i}, b_{1,i})^T \sim N(\mathbf{0}, \boldsymbol{D}) \end{aligned}$$

where $i \in \{1, ..., N\}$ is a group index while $j \in \{1, ..., n_i\}$ is an index for repeated measurements within group. We here assume that all b- and ε-variables are independent of each other.

(a) What is this model called?

Discuss the usefulness of such models.

(b) What is the *marginal* model for $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$?

What advantages do we have in the existence of an explicit expression of the marginal distribution for $\mathbf{Y}_i$ with respect to estimation?

(c) Discuss the main principles for REML estimation.

Discuss the advantages and disadvantages with maximum likelihood (ML) estimation compared to REML estimation. Specify in particular in which cases you would use the different methods.

Davidian and Giltinan (1995) describe a dataset for comparison of two types of soybeans. The dataset consists of 412 observations divided into 48 plots with 8-10 observations within each plot. In addition to weight (`weight`) and type soybean (`Variety`, two types) also the time the sample was taken (days after planting, `Time`) are given. The variable `Plot` specifies plot. In addition we define the variable `Time2` which is `Time` squared.

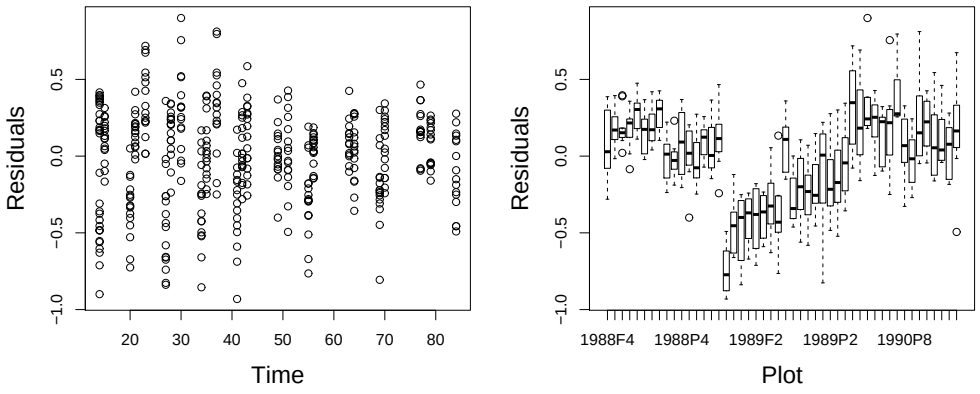
(Continued on page 3.)

(d) We will first look at a simple model where weight on log scale is used as response variable while `Variety`, `Time` and `Time2` are included as explanatory variables in addition to interaction between `Variety` and `Time`. We write the model in general as

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \varepsilon_{ij} \tag{M0}$$

where i specify plot while j specify repetition within plot.

The figure below shows boxplots of residuals grouped according to plots and a plot of residuals against `Time`. Comment on these plots and argue why a model similar to $(*)$ can be useful in this case.



(e) The model in the previous sub-exercise is now extended to two alternative models:

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij} \tag{M1}$$

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_{0,i} + b_{1,i}\text{Time}_{ij} + \varepsilon_{ij} \tag{M2}$$

where $b_i \sim N(0, d^2)$ in model M1 and $\boldsymbol{b}_i = (b_{0,i}, b_{1,i})^T \sim N(\boldsymbol{0}, \boldsymbol{D})$ in model M2. The log-likelihood values (with REML estimates inserted) for the three models are given below:

Model	M0	M1	M2
Loglik	-130.92	-13.42	6.16

Based on this, argue why model M2 is preferable.

(f) Below are the results based on a fit of model M2 performed by ML estimation. Discuss if there is any need for simplifying this model.

```
Linear mixed-effects model fit by maximum likelihood

Random effects:
 Formula: ~1 + Time | Plot
             StdDev      Corr
(Intercept) 0.373119650 (Intr)
```

(Continued on page 4.)

```
Time        0.002970683 -0.999
Residual    0.190066092

Fixed effects: log(weight) ~ Variety + Time + Time2 + Variety:Time
                  Value  Std.Error  DF   t-value p-value
(Intercept)   -5.202444 0.09214869 361 -56.45707  0.0000
VarietyP       0.478989 0.11677518  46   4.10181  0.0002
Time           0.204265 0.00236086 361  86.52125  0.0000
Time2         -0.001317 0.00002330 361 -56.52636  0.0000
VarietyP:Time -0.003079 0.00124599 361  -2.47102  0.0139
```

END

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100/4100 — Innføring i generaliserte lineære modeller.
Eksamensdag:	Mandag 5. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normal, χ^2 og t fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120.

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

De ulike delpunktene kan stort sett løses uavhengige av hverandre. Hvis du står fast på et punkt, gå derfor heller videre til neste punkt.

Oppgave 1

Vi skal i denne oppgaven se på følgende modell:

$$\begin{aligned} Y_{ij} &= \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \\ b_i &\sim N(0, \sigma_b^2) \end{aligned}$$

der $i \in \{1, ..., N\}$ er en gruppe-indeks mens $j \in \{1, ..., n_i\}$ er en indeks for repeterte målinger innen gruppe. Vi antar her alle tilfeldige variable er uavhengige av hverandre.

(a) Hva kalles denne modellen? (Bruk gjerne det engelske navnet)

Diskuter nytten av slike modeller.

(b) Hva blir den *marginale* modellen for $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$?

Hvilke fordeler har det at vi har et eksplisitt uttrykk for den marginale fordelingen til $\mathbf{Y}_i$ når det gjelder estimering?

Vi skal i den resterende delen av oppgaven se på et konkret datasett fra Havforskningsinstituttet i Bergen. Dette datasettet er en liten del av et større datasett som benyttes for å kartlegge bestander av fisk i Barentshavet. Vårt datasett vil begrense seg til observasjoner på torsk fra år 2000 innenfor et spesifikt område.
Følgende variable er tilgjengelige på et tilfeldig utvalg innenfor hver fangst (et utkast av tråd)

(Fortsettes på side 2.)

- `length` Lengde på fisk (i cm)
- `weight` Vekt på fisk (i gram)
- `age` Alder til fisk (i år)
- `haulsize` Størrelse på total fangst (i tonn)

Vi vil i det etterfølgende la i være en indeks for fangst (haul) mens j er indeks for en individuell fisk innen en fangst. Aldersvariabelen vil først bli brukt i neste oppgave.
Vi vil starte med å se på en modell

$$\log(\text{weight}_{ij}) = \beta_0 + \beta_1 \log(\text{length}_{ij}) + \beta_2 \log(\text{haulsize}_i) + b_i + \varepsilon_{ij}$$

der $b_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma^2)$ og alle tilfeldige effekter er uavhengige av hverandre. Nedenfor er en utskrift fra en tilpasning av denne modellen:

```
Linear mixed model fit by REML
Formula: log(weight) ~ log(length) + log(haulsize) + (1 | haul)
   Data: d.4
    AIC    BIC logLik deviance REMLdev
 -985.3 -961.4  497.6    -1012  -995.3
Random effects:
 Groups   Name        Variance  Std.Dev.
 haul     (Intercept) 0.0016294 0.040365
 Residual             0.0182412 0.135060
Number of obs: 885, groups: haul, 11

Fixed effects:
               Estimate Std. Error t value
(Intercept)   -4.028899   0.236925  -17.00
log(length)    2.873710   0.036685   78.33
log(haulsize) -0.002898   0.028023   -0.10

Correlation of Fixed Effects:
             (Intercept) log(length)
log(length)     -0.654
log(haulsize)   -0.732     -0.034
```

(c) Skriv opp estimatene til alle parametrene som inngår i modellen.

Hva blir korrelasjonen mellom to vekt-variable fra samme fangst (haul)?

(d) Vi vil i denne deloppgaven være interessert i

$$\theta = \exp\{\beta_0 + \beta_1 \log(66) + \beta_2 \log(0.46)\}$$

(Fortsettes på side 3.)

Du får her oppgitt at

$$\text{SE}\left(\hat{\beta}_0 + \hat{\beta}_1 \log(66) + \hat{\beta}_2 \log(0.46)\right) = 0.2009$$

der SE her står for standardfeil.

Forklar hvordan denne er blitt beregnet utifra opplysninger fra utskriften over (du behøver ikke å gjøre de faktiske utregninger).

Bruk dette til å lage et 95% konfidensintervall for θ.

(e) Anta nå vi ønsker å sammenlikne ulike modeller både med hensyn på hvilke tilfeldige effekter som bør være med *og* hvilke faste effekter som bør inkluderes. Skriv opp en generell strategi for å utføre modell-valg for slike modeller.

Oppgave 2

(a) Vis at den binomiske fordelingen er innenfor den eksponensielle klasse.

Hva er kanonisk link for den binomiske fordelingen innenfor GLM modellene?

Hva menes med kanonisk link, og hvilke fordeler har det å bruke kanonisk link?

Vi vil igjen se på data om fisk, men nå være interessert i alder. Som tidligere vil vi la indeks i stå for fangst (haul) mens j er indeks for individuell fisk innen fangst.
I utgangspunktet er alder en kategorisk variabel som varierer fra 3 til 13 år i dette datasettet. For å få det til å passe i vårt rammeverk, vil vi forenkle problemstillingen noe ved at vi vil definere

$$A_{ij} = \begin{cases} 1 & \text{hvis age}_{ij} > 9 \\ 0 & \text{ellers} \end{cases}$$

og bruke denne som responsvariabel.
Vi vil så se på følgende modell:

$$\begin{aligned} A_{ij} \sim & \text{Binom}(1, \pi_{ij}) \\ g(\pi_{ij}) = & \beta_0 + \beta_1 \log(\text{length}_{ij}) + \log(\text{haulsize}_i) \end{aligned}$$

der $g(\cdot)$ er en passende linkfunksjon og der alle A_{ij} er uavhengige.

(b) Tabellen nedenfor viser AIC verdier for 3 ulike link-funksjoner.

Link-funksjon	AIC
log	488.6953
probit	487.7484
cloglog	489.6841

(Fortsettes på side 4.)

Forklar hva AIC er og argumenter for hvorfor det er fornuftig å bruke et slikt kriterium (kontra andre typer tester vi har diskutert i kurset) i akkurat denne situasjonen.

Basert på disse verdiene, hvilken link-funksjon vil du foretrekke?

En utvidelse av modellen ovenfor er

$$\begin{aligned} A_{ij}|c_i &\sim \text{Binom}(1, \pi_{ij}) \\ g(\pi_{ij}) &= \beta_0 + \beta_1 \log(\text{length}_{ij}) + \beta_2 \log(\text{haulsize}_i) + c_i \\ c_i &\sim N(0, \sigma_c^2) \end{aligned}$$

Utskriften nedenfor svarer til denne modellen (link-funksjonen som er brukt her er ikke spesifisert og er ikke nødvendig å vite for å løse de følgende oppgaver, men er den optimale i forhold til AIC tabellen ovenfor).

```
Generalized linear mixed model fit by the Laplace approximation
Formula: A ~ log(length) + log(haulsize) + (1 | haul)
   Data: d.4
   AIC   BIC logLik deviance
 489.3 508.5 -240.7    481.3
Random effects:
 Groups Name        Variance Std.Dev.
 haul   (Intercept) 0.013427 0.11587
Number of obs: 885, groups: haul, 11

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -45.7253     3.7311 -12.255   <2e-16 ***
log(length)     9.6381     0.7910  12.184   <2e-16 ***
log(haulsize)   0.2781     0.1513   1.838   0.0661 .

Correlation of Fixed Effects:
            (Intercept) log(length)
log(length)    -0.965
log(haulsize)  -0.319      0.062
```

(c) Forklar hva som menes med at modellen er tilpasset med Laplace approksimasjon.

(d) Log-likelihood verdien for modellen uten tilfeldig effekt er -240.8. Bruk dette til å utføre en likelihood-ratio test på $H_0 : \sigma_c^2 = 0$. Beregn tilhørende P-verdi og konkluder.

(e) Havforskningsinstituttet mener at størrelsen på fangst er viktig for å modellere alder og lengde av fisk. Basert på utskriftene i både denne og foregående oppgave, hva er din mening om dette?

SLUTT

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære modeller
Eksamensdag:	Mandag 6. desember 2010
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over normalfordeling og χ^2-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK1120/STK2120

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

I vedlegg 1 er det et datasett fra 35 operasjoner der forekomsten av sår hals etter narkose er registrert. Her betrakter vi sår hals (`sore`) som respons, 0 svarer til nei og 1 til ja. Kovariatene er lengden på operasjonen i minutter (`duration`) og to typer (`type`) utstyr brukt til å holde luftveiene åpne under operasjonen.

Utskriften nedenfor er basert på en modell der responsen er binomisk fordelt, $Bin(m, \pi)$, der m er lik 1, altså det som kalles binære responser. Linkfunksjonen er logit link. I første omgang skal vi bare betrakte varighet som kovariat, dvs. prediktoren har formen

$$\eta = \beta_0 + \beta_1 x$$

der x er lengden på operasjonen.

```
Call:
glm(formula = sore ~ I(duration), family = binomial, data = sore)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0964  -0.7392   0.3020   0.8711   1.3753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.21358    0.99874  -2.216  0.02667 *
I(duration)  0.07038    0.02667   2.639  0.00831 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 33.651  on 33  degrees of freedom
AIC: 37.651
```

(Fortsettes på side 2.)

a) Forklar hvordan en generalisert lineær modell er definert, og hvorfor modellen ovenfor er av denne typen.

b) Estimer oddsforholdet for forekomst av sår hals mellom to operasjoner der den ene varer 30 og den andre 40 minutter. Angi også et 95% konfidensintervall for dette oddsforholdet.

c) Hva er den predikerte sannsynligheten for sår hals ved en operasjon som varer 40 minutter? Beregn også et 95% konfidensintervall. Her trenger du å vite at den estimerte korrelasjonen mellom $\hat{\beta}_0$ og $\hat{\beta}_1$ er -0.906.

d) I deviansanalyse-tabellen nedenfor finner du deviansen for modeller som også inneholder kovariaten `type` og et kvadratisk ledd i `duration`. Antallet frihetsgrader fjernet. Fyll ut de manglende tallene. Begrunn deretter at modellen vi så på punktene a)-c) er et rimelig valg. Du kan anta at den mest generelle modellen (`Model 4`) har en tilfredsstillende tilpasning.

```
Analysis of Deviance Table

Model 1: sore ~ 1
Model 2: sore ~ I(duration)
Model 3: sore ~ I(duration) + factor(type)
Model 4: sore ~ I(duration) + I(duration^2) + factor(type)
  Resid. Df Resid. Dev Df Deviance
1        ?     46.180
2        ?     33.651  ?   12.528
3        ?     30.138  ?    3.513
4        ?     30.133  ?    0.005
```

e) La y_i og $\hat{\pi}_i, 1 = 1, \ldots, 35$ være henholdsvis de observerte og de tilpassede responsverdiene. Vis at deviansen for binomiske modeller med binær respons kan skrives

$$-2\sum_{i=1}^{35}[\hat{\pi}_i \log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + \log(1-\hat{\pi}_i)].$$

Forklar omhyggelig hvorfor det medfører at deviansen er uegnet som føyningsmål i dette tilfellet.

Oppgave 2

Tabellen nedenfor er et berømt datasett fra en undersøkelse om sammenhengen mellom røyking og dødsfall på grunn av hjertesykdommer blant britiske leger. Antall dødsfall er respons og alder (`age`) og røyking (`smoker`) er kovariater. Vi lar alder være en numerisk variabel, og tilordner verdiene, eller skårene, 40, 50, 60, 70 og 80 til de fem aldersgruppene. Røyking er en faktor med to nivåer der 0 betegner ikke-røyker og 1 røyker. I tillegg er det registrert en variabel (`persyear`) som angir antall leveår i de ulike kategoriene.

(Fortsettes på side 3.)

Tabell 1: Dødelighet og røyking.

	Personår		Hjerterelatert dødsfall	
Alder	Ikke-røyker	Røyker	Ikke-røyker	Røyker
35-44	18793	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12633	28	186
75-84	1462	5317	31	102

Utskriften viser resultatet av en tilpasning av en modell der responsen har en Poisson-fordeling og det er benyttet en kanonisk log-link.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
    factor(smoker) + I(age):factor(smoker), family = poisson,
    data = coro)


Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.971e+01  1.253e+00 -15.734  < 2e-16 ***
I(age)                     3.565e-01  3.631e-02   9.819  < 2e-16 ***
I(age^2)                  -1.978e-03  2.736e-04  -7.228 4.89e-13 ***
factor(smoker)1            2.370e+00  6.559e-01   3.613 0.000303 ***
I(age):factor(smoker)1    -3.084e-02  9.699e-03  -3.180 0.001474 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 936.6589  on 9  degrees of freedom
Residual deviance:   1.6661  on 5  degrees of freedom
AIC: 66.734
```

a) Forklar hvorfor antagelsen om Poisson-fordelte responser er rimelig i denne situasjonen. Gi en eksplisitt beskrivelse av hvordan kovariatene inngår i modellen som er tilpasset i utskriften ovenfor. Kommenter resultatet.

b) Forklar hva offset er. Hvorfor er det rimelig å benytte offset i dette tilfellet?

c) Uttrykk betydningen av røyking for denne typen dødelighet ved relevante forhold mellom ratene (rate ratios). Beregn spesielt forholdene mellom ratene for røykere og ikke-røykere for leger som er 40 år og for leger som er 70 år. Diskuter resultatet.

(Fortsettes på side 4.)

Nedenfor finner du utskriften for tilpasning av en mer generell modell.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
    factor(smoker) + I(age):factor(smoker) + I(age^2):factor(smoker),
    family = poisson, data = coro)


Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.153e+01  3.197e+00  -6.736 1.63e-11 ***
I(age)                       4.148e-01  1.004e-01   4.130 3.62e-05 ***
I(age^2)                    -2.430e-03  7.739e-04  -3.140  0.00169 **
factor(smoker)1              4.445e+00  3.391e+00   1.311  0.18991
I(age):factor(smoker)1      -9.755e-02  1.069e-01  -0.912  0.36160
I(age^2):factor(smoker)1     5.196e-04  8.273e-04   0.628  0.52999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 936.6589  on 9  degrees of freedom
Residual deviance:   1.2623  on 4  degrees of freedom
AIC: 68.33

Number of Fisher Scoring iterations: 4
```

d) Vi ser at to siste estimatene som beskriver samspill mellom alder og røyking ikke er signifikante hver for seg. Utfør en Wald test for å teste hypotesen om at de tilsvarende koeffisientene er lik null samtidig, dvs den simultane hypotesen at begge er null. Matrisen nedenfor er den estimerte kovariansmatrisen for estimatorene til de to koeffesientene.

```
                            I(age):factor(smoker)1  I(age^2):factor(smoker)1
I(age):factor(smoker)1        1.143363e-02              -8.807653e-05
I(age^2):factor(smoker)1     -8.807653e-05               6.844424e-07
```

Den inverse kovariansmatrisen har diagonalelementer `10038.02` og `16768.5354e+04`. Elementene utenfor diagonalen er `12917.2852e+02`.

e) Forklar hvorfor den forventede og observerte informasjonsmatrisen blir like i modeller av den typen vi har sett på i denne oppgaven.

SLUTT

(Fortsettes på side 5.)

Vedlegg 1

	duration	type	sore
1	45	0	0
2	15	0	0
3	40	0	1
4	83	1	1
5	90	1	1
6	25	1	1
7	35	0	1
8	65	0	1
9	95	0	1
10	35	0	1
11	75	0	1
12	45	1	1
13	50	1	0
14	75	1	1
15	30	0	0
16	25	0	1
17	20	1	0
18	60	1	1
19	70	1	1
20	30	0	1
21	60	0	1
22	61	0	0
23	65	0	1
24	15	1	0
25	20	1	0
26	45	0	1
27	15	1	0
28	25	0	1
29	15	1	0
30	30	0	1
31	40	0	1
32	15	1	0
33	135	1	1
34	20	1	0
35	40	1	0

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — innføring i generaliserte lineære modeller.
Eksamensdag:	Tirsdag 15. desember 2009.
Tid for eksamen:	09.00 – 12.00.
Oppgavesettet er på 2 sider.	
Vedlegg:	Tabell over normalfordelingen.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

NB. Oppgavesettet består av to oppgaver.

Oppgave 1

(a) Vis at binomialfordelingen (n, π) tilhører den eksponensielle fordelingsklassen, der tettheten / punktsannsynligheten kan skrives på formen

$$f(y;\theta,\phi) = c(y;\phi)\exp((y\theta - a(\theta))/\phi).$$

Finn $a(\theta)$ og ϕ, og bruk dette til å vise de kjente formlene for forventning og varians i den binomiske fordelingen.

(b) Definer begrepet generalisert lineær modell (GLM). Hva menes med kanonisk link? Finn kanonisk link i en binær regresjonsmodell, dvs. en modell der responsen har en binær fordeling / Bernoullifordeling. Hvilke andre linkfunksjoner er vanlige for slike fordelinger?

(c) Definer begrepene mettet modell og devians. Se på en logistisk regresjonsmodell med regresjonsdel $\eta_i = \alpha + \beta x_i$, med binær responsvariabel y_i og med forklaringsvariabel x_i, $i = 1, 2, ..., n$. Hvordan vil du i denne modellen teste hypotesen $\mathrm{H}_0 : \beta = 0$ ved å gjøre bruk av devians-begrepet?

(d) Sett opp loglikelihoodfunksjonen for modellen i c). Vis at scorefunksjonen med hensyn til parametrene α og β har komponenter

$$\sum_{i=1}^{n}(y_i - \frac{e^{\eta_i}}{1+e^{\eta_i}}) \quad \text{og} \quad \sum_{i=1}^{n}(x_i y_i - \frac{x_i e^{\eta_i}}{1+e^{\eta_i}}).$$

Finn Fisherinformasjonsmatrisen. Fortell hvordan du ut fra disse størrelsene kan sette opp en iterasjonsprosedyre for maximum likelihood estimatoren av vektoren $(\alpha\ , \beta)^T$.

(Fortsettes på side 2.)

Oppgave 2

I en tysk spørreundersøkelse har man innhentet data om antall legebesøk siste 3 måneder (`numvisit`) og om ulike faktorer som kan ha betydning for legebesøk blant 1100 kvinner. Blant de innsamlede kovariatene skal vi bare se på en indikator for selvrapportert dårlig helse (`badh`) og alder (`age`) i år.

(a) Under er det gjengitt resultater fra en R-kjøring av en Poissonregresjon med log-link for antall legebesøk mot kovariatene `badh` og `age`. Spesifiser modellen matematisk og fortolk parameterne.

```
> M0<-glm(numvisit~age+badh,family=poisson)
> summary(M0)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9452  -2.0348  -0.8169   0.5191  12.5571

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.588731   0.064318   9.153  < 2e-16 ***
age         0.005556   0.001676   3.316 0.000914 ***
badh        1.140908   0.039858  28.625  < 2e-16 ***
---
Signif. codes:  0 ë***í 0.001 ë**í 0.01 ë*í 0.05 ë.í 0.1 ë í 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4779.4  on 1099  degrees of freedom
Residual deviance: 3975.3  on 1097  degrees of freedom
```

(b) Finn, basert på `R`-utskriften, estimater for rate-ratioer samt 95% konfidensintervall for legebesøk mellom

(i) kvinner med selvrapportert dårlig helse og selvrapportert god helse.

(ii) kvinner på 50 år og 40 år

Estimer også raten for legebesøk for en 40 årig kvinne med god helse. Hvilken ytterligere informasjon trenger du for å finne et konfidensintervall for (den teoretiske) raten?

(c) En mer generell modell tillater overspredning i forhold til Poissonmodellen i punkt a) via en spesifikasjon $\mathrm{Var}(Y) = \phi\mu$ der Y er responsen antall legebesøk og μ er forventningen til Y. En metode for å ta hensyn til overspredning består i å anta at $Y|Z$ er Poissonfordelt med forventning $Z\mu$ der Z er en latent variabel. Hvilken fordeling vil Y ha marginalt (ubetinget) hvis Z er gammafordelt (med forventning lik 1). Hvordan vil du begrunne svaret?

SLUTT

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK3100 — Innføring i generaliserte lineære modeller.
Eksamensdag:	Tirsdag 16. desember 2008.
Tid for eksamen:	09.00 – 12.00.
Oppgavesettet er på 3 sider.	
Vedlegg:	Tabell over normalfordelingen.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

NB. Oppgavesettet består av en oppgave med 8 punkter.

Oppgave 1.

(a) Vis at punktsannsynlighetene i Poissonfordelingen kan skrives på formen $f(y;\theta) = c(y)\exp(\theta y - a(\theta))$.

Finn sammenhengen mellom forventningen μ i Poissonfordelingen og parameteren θ.

Gi eksplisitte uttrykk for funksjonene $a(\theta)$ og $c(y)$.

(b) Anta at $Y_1, \ldots, Y_n$ er uavhengige og Poissonfordelte med forventning $\mu_i = \exp(\alpha + \beta x_i)$ der α og β er regresjonsparametre og x_i kjente forklaringsvariable. Påvis at denne modellen kommer inn under rammen for generaliserte lineære modeller.

(c) Sett opp log-likelihood for dataene i punkt (b) og vis at scorefunksjonen kan skrives

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

Finn også et uttrykk for forventet informasjonsmatrise.

(Fortsettes side 2.)

(d) Forklar hva en mettet modell er. Vis at maximum likelihood estimatene for $\mu_i, i = 1, \ldots, n$, i den mettede modellen blir $\tilde{\mu}_i = Y_i$.

(e) Finn på denne bakgrunn et uttrykk for deviansen ved en generalisert lineær modell med responser fra Poissonfordelingen.

Grei ut om tester som kan gjøres ved hjelp av devianser.

(f) Man velger å se bort fra antallet hendelser Y_i for hvert individ i og vil heller gjøre analysen basert på indikatorvariablen for at det var minst en skade, dvs. $Y_i' = I(Y_i > 0)$. La $\pi_i = \mathrm{P}(Y_i' = 1)$. Påvis at dette blir en generalisert lineær modell for binære data med samme lineære prediktor som i punkt (b) og linkfunksjon

$$g(\pi_i) = \log(-\log(1 - \pi_i)).$$

Hva kalles denne linkfunksjonen?

(g) Under er det gjengitt resultater fra en analyse av bilforsikringsdata der responsen er om forsikringstaker har rapportert en eller flere skader siste år. Modellen for dataene er en Poissonregresjonsmodell tilsvarende punkt (b), men dataene er analysert med binære responser tilsvarende den utledede modellen i punkt (f). Variablene som er brukt er bilførerens alder, kategorisert til 6 grupper (`agecat`), og bilens verdi i 10.000 dollar (`veh_value`) (brukt som en numerisk variabel slik at en bil som f.eks. er verdt 25.000 dollar er kodet som 2.5). Kun biler med verdi under 40.000 dollar er tatt med i analysen.

Beregn på basis av `R`-utskriften estimater for relativ forskjell i raten for skader mellom

(i) alderskategori 2 og alderskategori 1

(ii) alderskategori 5 og alderskategori 2

(iii) to biler verdt henholdsvis 25.000 og 5.000 dollar

(iv) to bilførere der den ene føreren er i alderskategori 2 og har en bil verdt 20.000 dollar, mens den andre føreren er i alderskategori 1 og har en bil verdt 10.000 dollar

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.79067    0.05636 -31.773  < 2e-16 ***
factor(agecat)2 -0.21697    0.05846  -3.711 0.000206 ***
factor(agecat)3 -0.25327    0.05674  -4.464 8.06e-06 ***
factor(agecat)4 -0.27294    0.05663  -4.820 1.44e-06 ***
factor(agecat)5 -0.51762    0.06396  -8.093 5.83e-16 ***
factor(agecat)6 -0.47167    0.07183  -6.567 5.14e-11 ***
veh_value        0.11656    0.01874   6.220 4.96e-10 ***
```

(Fortsettes side 3.)

(h) Finn dessuten 95% konfidensintervall for relativ endring i raten for skader mellom

(i) alderskategori 2 og alderskategori 1

(iv) to bilførere der den ene føreren er i alderskategori 2 og har en bil verdt 20.000 dollar, mens den andre føreren er i alderskategori 1 og har en bil verdt 10.000 dollar

For (iv) trenger du å vite at korrelasjonskoeffisienten mellom estimerte regresjonskoeffisienter svarende til aldersgruppe 2 og bilens verdi ble estimert til -0.0259.

SLUTT

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK3100/4100 — Innføring i generaliserte lineære modeller.
Eksamensdag:	Tirsdag 18. desember 2007.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

En eksponensiell klasse kan parametriseres ved tetthet / punktsannsynligheter på formen

$$f(y;\theta) = \exp(\theta y - c(\theta) + d(y))$$

(a) Vis at Poissonfordelingene kan bringes på denne formen.

(b) Vis at eksponensialfordelingene også kan skrives på denne formen.

(c) Finn forventning og varians i eksponensialfordelingen ved hjelp av karakteriseringen i punkt (b). Beskriv spesielt hvordan variansen avhenger av forventningen.

(d) En Paretofordelt Y variabel har kumulativ fordelingsfunksjon $F(y;\theta) = 1 - (1/y)^{\lambda}$ når $y > 1$. Vis at $V = \log(Y)$ er eksponensialfordelt med forventning $\mu = 1/\lambda$.

(e) Anta at $Y_i, i = 1, \ldots, n$ er uavhengige og Paretofordelte med parameter $\lambda_i = \exp(\alpha + \beta x_i)$ for kjente kovariater x_i. Begrunn at man kan bruke et statistikkprogram med en vanlig implementasjon av generaliserte lineære modeller (GLM) som f.eks. `R` til å estimere parametrene α og β.

(Fortsettes side 2.)

(f) Man kan noe mer generelt definere eksponensielle klasser ved at tetthetene kan skrives på formen

$$g(y;\theta) = \exp(\theta a(y) - c(\theta) + d(y))$$

Vis at med denne definisjonen tilhører Paretofordelingene den eksponensielle fordelingsklasse.

(g) Anta at Y har tetthet på formen $g(y;\theta) = \exp(\theta a(y) - c(\theta) + d(y))$. Utled et uttrykk for tettheten til $V = a(Y)$. Angi spesielt sammenhengen med parametriseringen $f(v;\theta)$ fra innledningen til denne oppgaven.

Oppgave 2.

I denne oppgaven skal vi se på risikoen for dødelighet i den såkalte "postneonatale perioden", fra 28. levedag til ett-årsdag. Vi skal bare se på dødelighet av SIDS (Sudden infant death syndrome, plutselig spebarnsdød) og vi modellerer sannsynligheten for SIDS-død i perioden, gitt at barna ikke dør av en annen årsak, ved logistisk regresjon.
Vi skal spesielt se på hvordan SIDS-dødeligheten avhenger av fødselsår, kjønn og fødselsvekt. Fødselsår (`kohort`) er angitt som en kategorisk variabel med 5 nivåer der nivå 1 angir 1967-1974, nivå 2 1975-1979, nivå 3 1980-1984, nivå 4 1985-1989 og nivå 5 1990-1995. Kjønn (`kjonn`) er kodet som 1 for gutter og 2 for jenter. Fødselsvekt (`vekt`) benyttes som kontinuerlig variabel angitt i kilo.

(a) Under er det gjengitt en `R`-utskrift av en devianstabell for dataene hvor endel størrelser er byttet ut med "?". Fyll ut verdiene som er byttet ut og fortolk resultatene. Når det gjelder p-verdier er det nok å angi om vi har statistisk signifikante effekter eller ikke.

Gi en begrunnelse for at signifikanstesting kan gjennomføres på bakgrunn av deviansanalysetabellen.

```
Analysis of Deviance Table
Model: binomial, link: logit
```

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			570	1101.92	
vekt	?	?	569	842.33	?
factor(kohort)	?	?	?	527.74	?
kjonn	?	?	?	434.93	?
vekt:factor(kohort)	?	?	?	428.56	?
vekt:kjonn	?	?	?	428.37	?
factor(kohort):kjonn	?	?	?	413.05	0.0041
vekt:factor(kohort):kjonn	?	?	?	407.80	?

(Fortsettes side 3.)

(b) Under er det gjengitt en `R`-utskrift hvor det bare er tatt hensyn til hovedeffektene av kovariatene `vekt, kjonn` og `kohort`. For `kohort` er det benyttet en hjørnepunktparametrisering med nivå 1967-1974 som referanse. Fortolk resultatene ved hjelp av odds-ratio-estimater avledet fra tabellen.

Beregn også et 95% konfidensintervall for odds-ratioen svarende til fødselsvekt.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.37607    0.15833 -27.639  < 2e-16 ***
vekt            -0.67110    0.03758 -17.859  < 2e-16 ***
kjonn           -0.47371    0.04981  -9.511  < 2e-16 ***
factor(kohort)2  0.56224    0.08629   6.515 7.25e-11 ***
factor(kohort)3  0.90941    0.08105  11.220  < 2e-16 ***
factor(kohort)4  1.07958    0.07743  13.943  < 2e-16 ***
factor(kohort)5  0.11049    0.08958   1.233    0.217
---
```

(c) Vi har så langt ignorert dødelighet av andre årsaker enn SIDS. Anta nå at det totalt er J ulike dødsårsaker og at vi har en multinomisk regresjonsmodell med $J+1$ utfall (inkludert de overlevende) med π_{ij} lik sannsynligheten for at individ i dør av årsak $j = 1, \ldots, J$ og π_{i0} lik sannsynlighet for at individ i overlever. Med kovariatvektorer x_i angis modellen ved

$$\pi_{ij} = \frac{\exp(\beta_j' x_i)}{1+\sum_{k=1}^{J} \exp(\beta_k' x_i)} \text{ for } j = 1, \ldots, J$$

$$\pi_{i0} = \frac{1}{1+\sum_{k=1}^{J} \exp(\beta_k' x_i)}$$

der β_j er vektorer av regresjonsparametre. Vi angir SIDS som årsak $j = 1$.

Vis at vi da har en vanlig logistisk regresjonsmodell for å dø av SIDS gitt at vi (som i punkt (a) og (b)) bare ser på de som dør av SIDS og de som overlever. Angi spesielt parametrene i denne logistiske regresjonsmodellen.

Diskuter ulemper og fordeler med å analysere dataene med logistisk regresjon framfor den fulle multinomiske regresjonsmodellen.

SLUTT

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK3100/STK4100 — Introduction to Generalized Linear Models
SKETCH OF SOLUTION

Day of examination: Thursday 14th December 2023

This problem set consists of 0 pages.

Appendices:

Permitted aids:

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

Problem 1

a

We can write (1) as

$$P(Y=y) = \exp\{\log P(Y=y)\} = \exp\left\{\log\binom{n}{ny} + ny\log\pi + (n-ny)\log(1-\pi)\right\}$$
$$= \exp\left\{\log\binom{n}{ny} + ny\log\frac{\pi}{1-\pi} + n\log(1-\pi)\right\}$$

which is identical to $f(y;\theta,\phi)$ with $\theta = \log\frac{\pi}{1-\pi}$, which means that $\pi = \frac{e^\theta}{1+e^\theta}$. Furthermore, $a(\phi) = 1/n$, $b(\theta) = -\log(1-\pi) = \log\left(1+e^\theta\right)$ and $c(y,\phi) = \log\binom{n}{ny}$.

b

(i)

$$E(Y) = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \pi$$

(ii)

$$\mathrm{Var}(Y) = a(\phi)b''(\theta) = \frac{1}{n}\frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \frac{1}{n}\frac{e^\theta}{(1+e^\theta)^2} = \frac{1}{n}\pi(1-\pi)$$

c

(i) The canonical link function is defined to be the natural parameter in the exponential dispersion distribution formulation., i.e. $g(\pi_i) = \theta_i = \log\frac{\pi_i}{1-\pi_i}$.

(ii) This GLM is called a logistic regression model.

(Continued on page 2.)

(iii) The likelihood equations for this GLM are (using the formula given in the appendix)

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1$$

with $x_{i0} = 1$ and $x_{i1} = x_i$. Using $\mu_i = b'(\theta_i) = \pi_i = \frac{e^\theta}{1+e^\theta}$, and hence $\frac{\partial \mu_i}{\partial \eta_i} = b''(\theta_i) = \frac{e_i^\theta}{\left(1+e_i^\theta\right)^2} = \pi_i(1-\pi_i)$, and $\text{var}(Y_i) = \frac{1}{n_i}\pi_i(1-\pi_i)$, we get the following two equations

$$\text{For } j = 0: \quad \sum_{i=1}^{N} \frac{(y_i - \pi_i)}{\frac{1}{n_i}\pi_i(1-\pi_i)} \pi_i(1-\pi_i) = \sum_{i=1}^{N} n_i(y_i - \pi_i) = 0$$

$$\text{For } j = 1: \quad \sum_{i=1}^{N} \frac{(y_i - \pi_i)x_i}{\frac{1}{n_i}\pi_i(1-\pi_i)} \pi_i(1-\pi_i) = \sum_{i=1}^{N} n_i(y_i - \pi_i)x_i = 0$$

d

We have (using the formula given in the appendix)

$$w_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\text{var}(Y_i)} = \frac{(\pi_i(1-\pi_i))^2}{\frac{1}{n_i}\pi_i(1-\pi_i)} = n_i\pi_i(1-\pi_i)$$

e

(i) The deviance for this GLM is (using the formula given in the appendix)

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{N} \omega_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]$$

where $\omega_i = n_i$, $\tilde{\theta}_i$ is the ML estimate of θ under the saturated model and $\hat{\theta}_i$ is the ML estimate of θ under the actual model. Using that $\theta_i = \log \frac{\pi_i}{1-\pi_i}$, $b(\theta_i) = -\log(1-\pi_i)$ and $\tilde{\pi}_i = y_i$, we get

$$\begin{aligned} D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{N} n_i \left[y_i \left(\log \frac{\tilde{\pi}_i}{1-\tilde{\pi}_i} - \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} \right) + \log(1-\tilde{\pi}_i) - \log(1-\hat{\pi}_i) \right] \\ &= 2\sum_{i=1}^{N} \left[n_i y_i \log \left(\frac{n_i y_i}{n_i \hat{\pi}_i} \right) + (n_i - n_i y_i) \log \left(\frac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i} \right) \right] \end{aligned}$$

(ii) For ungrouped data, the deviance can be used to compare two nested models. Consider two models M_0 (with p_0 parameters) and M_1 (with p_1 parameters), with M_0 nested in M_1. If the null hypothesis is that M_0 holds, then the likelihood ratio statistics becomes (for GLMs with $a(\phi) = 1/\omega_i$, which is the case for binomial GLMs) $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1)$, which is approximately chi-squared distributed with $p_0 - p_1$ degrees of freedom.

(Continued on page 3.)

Problem 2

a

(i) The model used is a logistic regression model, which assumes that the binary response variable Y_i is $\text{bin}(1, \pi_i)$ distributed, $i = 1, \ldots,$ and that the Y_i's are independent. Furthermore, it is assumed that there is a linear predictor $\eta_i = \sum_{j=1}^{8} \beta_j x_{ij}$, that is linked to the mean $\mu_i = E(Y_i) = \pi_i$ through the canonical link function $\eta_i = g(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$. Here, for $i = 1, \ldots N$, $x_{i1} = 1$ represents the intercept, x_{i2} represents gender, x_{i3} represents age, x_{i4} represents indicator of smoking status, x_{i5} represents average number of cigarettes, x_{i6} represents cholesterol level, x_{i7} represents systolic blood pressure and x_{i8} represents glucose level.

(ii) The estimate $\hat{\beta}_2$ belonging to the explanatory variable `male` can be interpreted by considering $\exp(\hat{\beta}_2) = \exp(0.55) = 1.73$, which is the relative effect on the odds for a male versus a female, when they have identical values for all other explanatory variables. That is, given this model, a male has an estimated 73% higher odds of experiencing CHD during a 10 year period than a female, when the values for all other explanatory variables are the same.

b

(i) This output indicates that `currentSmoker` should be dropped from the model because the p-value of the likelihood ratiod-test for a null-model without this vs the original model is very high (0.76), and the highest of all the corresponding tests for dropping each of the other explanatory variables. Also the AIC for a model without this explanatory variable is lower than for the original model, and lower than all other models where one of the current explanatory variables is dropped.

(ii) A possible reason why smoking status `currentSmoker` does not seem to be significant in this model is that we also have the `cigsPerDay` variable in the model, which is 0 for non-smokers, and > 0 for smokers, and hence is correlated with `currentSmoker`, making `currentSmoker` redundant.

(Continued on page 4.)

c

(i) The numbers are filled in below

```
Analysis of Deviance Table

Model 1: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose
Model 2: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose +
    totChol:glucose
Model 3: TenYearCHD ~ male + age + cigsPerDay + totChol + sysBP + glucose +
    totChol:glucose + totChol:sysBP
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3808     2908.3
2      3807     2904.8  1   3.5580  0.05926 .
3      3806     2903.8  1   0.9741  0.32366
```

(ii) The p-value comparing models 1 and 2 is quite low, but above 5%, and it seems best to choose the model with the fewest parameters.The test comparing models 2 and 3 favours model 2 over model 3, but since model 1 is favoured over model 2, we conclude that model 1 seems from the given output to have the best fit of the three models.

Problem 3

a

(i) We have

$$E(Y_{ij}) = E[E(Y_{ij}|u_i)] = E[\exp(\beta_0 + \beta_1 x_{ij} + u_i)] = \exp(\beta_0 + \beta_1 x_{ij})\, E(\exp(u_i))$$

(ii) We have $u_i \sim N(0, \sigma_u^2$. Using the hint that for this distribution $M(t) = \exp(\sigma_u^2 t^2/2)$, and that by definition the moment generating function is $M(t) = E(\exp(u_i t))$, we have that

$$E(\exp(u_i)) = \exp(\sigma_u^2/2)$$

(iii) Since for the log-link Poisson GLMM we marginally get $E(Y_{ij}) = \exp\left(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2\right)$, while for the marginal model $E(Y_{ij}) = \exp(\beta_0 + \beta_1 x_{ij})$, the effect of the explanatory variable on the mean of Y_{ij} is the same for the log-link Poisson GLMM and the marginal model, while the intercept differs by $\sigma_u^2/2$.

b

(i) Since we know from a) that the effect of the explanatory variable on the mean of Y_{ij} is the same for the log-link Poisson GLMM and the marginal model, we can get the number behind the question mark from the output for fitting the GLMM, that is 0.3632.

(ii) The "Naive S.E." results from the assuming that the "working covariance matrix" is correctly specified, while the "Robust S.E." allows for the "working covariance matrix" to be misspecified.

(Continued on page 5.)

(iii) The sandwich estimator.

(Continued on page 6.)

APPENDIX: Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \ldots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \ldots, \eta_n$ are given by $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and

the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right), \quad \text{approximately}$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \omega_i \left[y_i \left(\tilde{\theta}_i - \widehat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ for $i = 1, \ldots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots \epsilon_{id})^{\mathrm{T}} \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \,|\, \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK3100/STK4100 — Introduction to Generalized Linear Models
SKETCH OF SOLUTION

Day of examination: Thursday 8th December 2022

This problem set consists of 5 pages.

Appendices:

Permitted aids:

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

Problem 1

a

We can write (1) as

$$P(Y=y) = \exp\{\log P(Y=y)\} = \exp\{y\log\mu - \log y! - \mu\},$$

which is identical to (2) with $\theta = \log\mu$, $b(\theta) = \mu = \exp(\theta)$, $a(\phi) = 1$ and $c(y,\phi) = -\log y!$.

b

The definition of a generalized linear model (GLM) for $Y_i \sim \text{Poisson}(\mu_i)$ and covariates $x_{ij}, i = 1,\dots n, j = 1,\dots,p$, with $x_{i1} = 1$ to represent the intercept, using the canonical link function

- $Y_1,\dots,Y_N$ are independent with $Y_i \sim \text{Poisson}(\mu_i)$ (i.e. pmf of the form (**??**))
- For each Y_i with covariates $x_{ij}, i = 1,\dots n, j = 1,\dots,p$, we have the linear predictor $\eta_i = \sum_{j=1}^{p}\beta_j x_{ij} = \boldsymbol{x}_i\boldsymbol{\beta}$, for $\boldsymbol{x}_i = (x_{i1},\dots,x_{ip})$ and $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^T$
- For each Y_i, the mean $\mu_i = E(Y_i)$ is linked to the linear predictor by the link function $g(\mu_i) = \eta_i$. Here we are using the canonical link function $g(\mu_i) = \theta_i = \log\mu$.

c

The log-likelihood function is

$$L(\boldsymbol{\mu};\mathbf{y}) = \log\left\{\prod_{i=1}^{n}\exp\{y_i\log\mu_i - \mu_i - \log y_i!\}\right\} = \sum_{i=1}^{n}\{y_i\log\mu_i - \mu_i - \log y_i!\} \tag{1}$$

(Continued on page 2.)

The saturated model is the most general model a separate parameter μ_i for each subject i, hence without any restrictions on μ_i. For this model the maximum likelihood fit $\tilde{\mu}_i = y_i$, because $\frac{\partial L(\boldsymbol{\mu};\mathbf{y})}{\partial \mu_i} = \frac{y_i}{\mu_i} - 1$ is $= 0$ for $\tilde{\mu}_i = y_i$. It is useful as a baseline for assessing the quality of fit for other models, as it is the model out of all possible models that achieves the maximum achievable value of the log likelihood $L(\mathbf{y};\mathbf{y})$. It is however not a very useful model for estimation of the underlying truth nor for prediction of new observations, e.g. it overfits the data and does not smooth. The maximum log likelihood for the saturated model is

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

d

The deviance $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ for a Poisson GLM is (because $a\phi = \phi = 1$)

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = -2\left[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})\right] = -2\left[\sum_{i=1}^{n}\{y_i \log \hat{\mu}_i - \hat{\mu}_i - \log y_i!\} - \sum_{i=1}^{n}\{y_i \log y_i - y_i - \log y_i!\}\right]$$
$$= 2\sum_{i=1}^{n}\left\{y_i \log \frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i\right\}$$

This deviance may be used as a test statistic to compare nested Poisson GLMs by considering the difference in the deviance for the smaller model and the bigger model. To explain this, assume that the Poisson GLM model M_1 with p_1 parameters holds, and that M_0 with $p_0 < p_1$ parameters is nested in model M_1 (i.e. has the same distribution and link function as M_1, here Poisson with log-link, and the same linear predictor as M_1, but with $p_1 - p_0$ of the β_j's of M_1 set to zero). Let $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ be the maximum likelihood estimates of μ under M_0 and M_1 respectively. For testing the null hypothesis that M_0 holds, we have the likelihood ratio statistics

$$G^2(M_0 \mid M_1) = -2\left(L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})\right) = -2\left(L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\mathbf{y};\mathbf{y})\right) - \left(-2\left(L(\hat{\boldsymbol{\mu}}_1;\mathbf{y}) - L(\mathbf{y};\mathbf{y})\right)\right)$$
$$= D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1)$$

which is approximately $\mathcal{X}^2_{p_1-p_0}$ distributed if M_0 holds.

Problem 2

In this problem we assume that Y comes from a negative binomial distribution. Here we let the pmf of the negative binomial distribution take the form

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

We will assume that $k > 0$ is a given constant, and consider the random variable $Y^* = Y/k$. Then $P(Y^* = y^*) = P(Y = ky^*)$ for $y^* = 0, \frac{1}{k}, \frac{2}{k}, \ldots$, so Y^* has pmf

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

(Continued on page 3.)

a

We can write (2) as

$$\begin{aligned} p^*(y^*;\mu,k) &= \exp\log p^*(y^*;\mu,k) \\ &= \exp\left\{\log\Gamma(y+k) - \log\Gamma(k) - \log\Gamma(ky^*+1) + ky^*\log\left(\frac{\mu}{\mu+k}\right) + k\log\left(\frac{k}{\mu+k}\right)\right\} \\ &= \exp\left\{\frac{y^*\log\left(\frac{\mu}{\mu+k}\right) + \log\left(\frac{k}{\mu+k}\right)}{\frac{1}{k}} + c(y,\phi)\right\} = \exp\left\{\left(\theta y^* - b(\theta)\right)/a(\phi) + c(y^*,\phi)\right\} \end{aligned}$$

where $\theta = \log\left(\frac{\mu}{\mu+k}\right)$, $b(\theta) = -\log\left(\frac{k}{\mu+k}\right) = -\log\left(1-\frac{\mu}{\mu+k}\right) = -\log\left(1-e^\theta\right)$ and $a(\phi) = 1/k$ (and $c(y^*,\phi) = \log\Gamma(y+k) - \log\Gamma(k) - \log\Gamma(ky^*+1)$).

b

Mean of Y^*

$$E[Y^*] = b'(\theta) = \frac{e^\theta}{1-e^\theta} = \frac{\frac{\mu}{\mu+k}}{1-\frac{\mu}{\mu+k}} = \frac{\mu}{\mu+k-\mu} = \frac{\mu}{k}$$

Variance of Y^*

$$\mathrm{Var}[Y^*] = b''(\theta)\cdot a(\phi) = \frac{e^\theta(1-e^\theta) - e^\theta(-e^\theta)}{(1-e^\theta)^2}\cdot\frac{1}{k} = \frac{e^\theta}{k(1-e^\theta)^2} = \frac{\frac{\mu}{\mu+k}}{k(1-\frac{\mu}{\mu+k})^2} = \frac{\mu(\mu+k)}{k^3}$$

which gives us

$$E[Y] = E[kY^*] = kE[Y^*] = k\frac{\mu}{k} = \mu$$

and

$$\mathrm{Var}[Y] = \mathrm{Var}[kY^*] = k^2\mathrm{Var}[Y^*] = k^2\frac{\mu(\mu+k)}{k^3} = \mu + \frac{\mu^2}{k}$$

c

If Y is negative-binomially distributed we have $\mathrm{Var}[Y]/E[Y] = 1+\mu/k$, and if Y is Poisson distributed we have $\mathrm{Var}[Y]/E[Y] = 1$. Hence if $k < \infty$, or $\gamma = 1/k > 0$, the variance is larger than the mean for the negative binomial distribution, while they are equal for the Poisson distribution. When you have count data for which it is reasonable to assume equal mean and variance, the Poisson is a good model, while if the variance is assumed to be larger than the mean, there is overdispersion and the negative binomial is a better model than the Poisson.

Problem 3

a

The model behind the analysis is a Poisson GLM with log link. Hence

- The response variables $Y_1,\ldots,Y_N$ are assumed to be independent with $Y_i \sim \mathrm{Poisson}(\mu_i)$ (i.e. pmf of the form (**??**))

(Continued on page 4.)

- For each Y_i with covariates $x_{ij}, i = 1, \dots n, j = 1, \dots, 5$, we have the linear predictor $\eta_i = \sum_{j=1}^{5} \beta_j x_{ij} = \boldsymbol{x}_i \boldsymbol{\beta}$, for $\boldsymbol{x}_i = (x_{i1}, \dots, x_{i5})$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_5)^T$. We have here
  - $x_{i1} = 1, \forall i$ to represent the intercept
  - $x_{i2} = 1$ if `procedure` of patient i is CABG, and $x_{i2} = 0$ if PTCA
  - $x_{i3} = 1$ if patient i is male, and $x_{i3} = 0$ if female
  - $x_{i4} = 1$ if patient i was admitted as an emergency, and $x_{i4} = 0$ if the procedure was pre-planned
  - $x_{i5} = 1$ if the age of patient i is over 75, and $x_{i5} = 0$ if it is less than or equal to 75
- For each Y_i, the mean $\mu_i = E(Y_i)$ is linked to the linear predictor by the link function $g(\mu_i) = \eta_i$. Here we are using the canonical link function $g(\mu_i) = \theta_i = \log \mu$, and hence $\mu_i = E(Y_i) = \exp\left\{\sum_{j=1}^{5} \beta_j x_{ij}\right\}$

b

The p-value of the likelihood ratio test described in Problem 1d for comparing the model with interaction (M_1) to the model without interactions (M_0) is reported to be $< 2.2e-16$, which is very small, which means that there is reason to reject M_0 and conclude that model M_1 with interactions is to be preferred. It can be commented that this is supported by the fact that the AIC is lower for M_1 than M_0.

To describe the effects of `procedure` and `admit` on the estimated mean length of stay, we consider a female patient ($x_{i3} = 0$) of age less than or equal to 75 ($x_{i5} = 0$). The estimated mean length of stay for the different combinations of `procedure` and `admit` are then

- `procedure` of patient i is PTCA ($x_{i2} = 0$) and the procedure was pre-planned ($x_{i4} = 0$):

$$\exp\left\{\sum_{j=1}^{5} \hat{\beta}_j x_{ij}\right\} = \exp\left\{\hat{\beta}_1\right\} = \exp\{1.23851\} \approx 3.5$$

- `procedure` of patient i is CABG ($x_{i2} = 1$) and the procedure was pre-planned ($x_{i4} = 0$):

$$\exp\left\{\sum_{j=1}^{5} \hat{\beta}_j x_{ij}\right\} = \exp\left\{\hat{\beta}_1 + \hat{\beta}_2\right\} = \exp\{1.23851 + 1.24765\} \approx 12$$

- `procedure` of patient i is PTCA ($x_{i2} = 0$) and the procedure was an emergency ($x_{i4} = 1$):

$$\exp\left\{\sum_{j=1}^{5} \hat{\beta}_j x_{ij}\right\} = \exp\left\{\hat{\beta}_1 + \hat{\beta}_4\right\} = \exp\{1.23851 + 0.61606\} \approx 6.4$$

(Continued on page 5.)

- `procedure` of patient i is CABG ($x_{i2} = 1$) and the procedure was an emergency ($x_{i4} = 1$):

$$\exp\left\{\sum_{j=1}^{5}\hat{\beta}_j x_{ij}\right\} = \exp\left\{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_6\right\} = \exp\left\{1.23851 + 1.24765 + 0.61606 - 0.39658\right\}$$
$$\approx 15$$

c

The p-value of the test is (close to) 0 and hence the null hypothesis of $\gamma = 0$ should be rejected. This implies that there is overdispersion and the negative binomial is to be preferred over the Poisson. This is supported by the fact that the AIC value for the fit with the negative binomial (19857) is lower than the one for the Poisson (22184).

Problem 4

a

The estimated β_j's are the same, which is not surprising, since they are estimated by the exact same equations. Their estimated standard errors are however different, smaller for the Poisson GLM. This is because the ones for the Poisson are based on an inadequate assumption for the variance function (that it is equal to the mean), while the quasi-likelihood allows for this to be multiplied by ϕ. Hence the estimated standard errors of the estimated β_j's of the quasi-likelihood approach are more robust and reliable.

b

The likelihood equations for a GLM with variance function $\text{var}(Y_i) = v^*(\mu_i)$ are

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i)x_{ij}}{v^*(\mu_i)}\frac{\partial\mu_i}{\partial\eta_i} = 0, \qquad j = 1, \ldots, p$$

The only dependence on the assumed distributon is here through how the variance function $v^*(\mu_i)$ depends on the mean μ_i.

We can allow for over-dispersion and instead assume $\text{var}(Y_i) = v(\mu_i) = \phi v^*(\mu_i)$. Then we have the quasi-likelihood equations

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i)x_{ij}}{\phi v^*(\mu_i)}\frac{\partial\mu_i}{\partial\eta_i} = 0, \qquad j = 1, \ldots, p$$

that are not dependent on any particular distribution, only that $\text{var}(Y_i) = v(\mu_i) = \phi v^*(\mu_i)$ and $g(\mu_i) = g(E[Y_i]) = \eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ for a link function g. ϕ can be estimated

These equations give estimates of $\boldsymbol{\beta}$ that are (as mentioned in a)) the same as obtained fro the ML estimates for the original GLM (because ϕ cancels from the quasi-likelihood equations), and they are approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix equal to the covariance matrix obtained fro the ML estimates for the original GLM multiplied by ϕ, which can be estimated.

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to Generalized Linear Models SKETCH OF SOLUTION
Day of examination:	Monday 6th December 2021
Examination hours:	15.00 – 19.00
This problem set consists of 5 pages.	
Appendices:	Formulas in STK3100/4100
Permitted aids:	Approved calculator

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

Problem 1

a

We have

$$\frac{\partial \log f(Y;\theta,\phi)}{\partial \theta} = \left(Y - b'(\theta)\right)/a(\phi),$$

and hence combined with the hint

$$E\left[\left(Y - b'(\theta)\right)/a(\phi)\right] = 0 \quad \Rightarrow \quad E[Y] = b'(\theta).$$

Furthermore we have

$$\frac{\partial^2 \log f(Y;\theta,\phi)}{\partial \theta^2} = -b''(\theta)/a(\phi),$$

which combined with the result $E[Y] = b'(\theta)$ shown above and the hint gives

$$\begin{aligned} -E\left[-b''(\theta)/a(\phi)\right] &= E\left[\left(\left(Y - b'(\theta)\right)/a(\phi)\right)^2\right] \\ &\Downarrow \\ b''(\theta)/a(\phi) &= E\left[\left(Y - b'(\theta)\right)^2\right]/a(\phi)^2 \\ &\Downarrow \\ E\left[\left(Y - E[Y]\right)^2\right] &= \mathrm{Var}[Y] = b''(\theta)a(\phi) \end{aligned}$$

b

We can write

$$P(Y = y) = \exp\left\{\log P(Y = y)\right\} = \exp\left\{y \log \mu - \log y! - \mu\right\}$$

which is identical to (1) with $\theta = \log \mu$, $b(\theta) = \mu = \exp(\theta)$, $a(\phi) = 1$ and $c(y, \phi) = -\log y!$.

(Continued on page 2.)

c

From the results proven in **a** we get

$$E[Y] = b'(\theta) = \exp(\theta) = \mu$$
$$\text{Var}[Y] = b''(\theta)a(\phi) = \exp(\theta) = \mu$$

d

The log-likelihood function is

$$L(\beta_0, \beta_1) = \log\left\{\prod_{i=1}^{n} \exp\{Y_i \log\mu_i - \log Y_i! - \mu_i\}\right\} = \sum_{i=1}^{n}\{Y_i \log\mu_i - \log Y_i! - \mu_i\}$$
$$= \sum_{i=1}^{n}\{Y_i\eta_i - e^{\eta_i} - \log Y_i!\} \quad (1)$$

From (1) we get for $j = 0, 1$, with $x_{i0} = 1$ and $x_{i1} = x_i$,

$$\frac{\partial L(\beta_0,\beta_1)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\{Y_i\eta_i - e^{\eta_i} - \log Y_i!\}}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = \sum_{i=1}^{n}(Y_i - e^{\eta_i})\frac{\partial\eta_i}{\partial\beta_j}$$
$$= \sum_{i=1}^{n}(Y_i - \mu_i)\,x_{ij} = 0$$

which implies $\sum_{i=1}^{n}(Y_i - \mu_i) = 0$ and $\sum_{i=1}^{n}(Y_i - \mu_i)x_i = 0$.

It is also ok to find $\hat{\beta}_0$ and $\hat{\beta}_1$ from the likelihood equations for a GLM given in the appendix.

Problem 2

a

The model behind the fit reported in **a** is a logistic regression model. For $i = 1, \ldots, n$ ($n = 2702$), let Y_i denote the response variable, where $Y_i = 1$ if there was objection against patent i, and $Y_i = o$ if there was no objection against patent i, and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7})$ denote the vector of explanatory variables, where $x_{i1} = 1$ for the intercept and

- x_{i2} = grant year
- x_{i3} = number of citations for the patent
- $x_{i4} = 1$ if US twin patent exists, otherwise $x_{i4} = 0$
- $x_{i5} = 1$ if patent holder is from the US, otherwise $x_{i5} = 0$
- $x_{i6} = 1$ if patent holder is from Germany, Switzerland or Great Britain, otherwise $x_{i6} = 0$
- x_{i7} = number of designated countries for the patent

(Continued on page 3.)

Let $\pi_i = P(Y_i = 1)$. The model assumes that $Y_1, \ldots, Y_n$ are independent and, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_7)^T$, that

$$\text{logit}(\pi_i) = \frac{\pi_i}{1-\pi_i} = \eta_i = \mathbf{x}_i\boldsymbol{\beta} = \sum_{j=1}^{7} \beta_j x_{ij},$$

which is equivalent to

$$\pi_i = \frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}$$

The estimate $\hat{\beta}_5$ for `patus` can be interpreted by considering two different patents i and k, that have the same values of all the explanatory variables except `patus`, i.e. $x_{ij} = x_{kj}$ for $j = 1, 2, 3, 4, 6, 7$ and $x_{i5} \neq x_{k5}$. In words they share grant year, number of citations for the patent, either both patent holders have a US twin patent or not, neither of the patent holders are from Germany, Switzerland or Great Britain, and they have the same number of designated countries, but the holder of patent i is from the US, while the holder of patent k is not. Then the odds ratio for the two patents is is

$$\begin{aligned}
&\frac{P(Y_i = 1)/\left[1 - P(Y_i = 1)\right]}{P(Y_k = 1)/\left[1 - P(Y_k = 1)\right]} = \frac{\pi_i/\left[1-\pi_i\right]}{\pi_k/\left[1-\pi_k\right]} \\
&= \frac{\frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)} \Big/ \left[1 - \frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}\right]}{\frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{kj}\right)}{1+\exp\left(\sum_{j=1}^{7} \beta_j x_{kj}\right)} \Big/ \left[1 - \frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{kj}\right)}{1+\exp\left(\sum_{j=1}^{7} \beta_j x_{kj}\right)}\right]} = \frac{\exp\left(\sum_{j=1}^{7} \beta_j x_{ij}\right)}{\exp\left(\sum_{j=1}^{7} \beta_j x_{kj}\right)} = e^{\beta_5}
\end{aligned}$$

Hence the estimated odds ratio for the two patents is $e^{\hat{\beta}_5} = e^{-0.43685} = 0.646$, which means that $\hat{\beta}_5$ tells us that we estimate that the odds of experiencing opposition against a patent is reduced by 35.4% when the patent holder is from the US.

b

The missing numbers are filled in in the table below. They are found in the following way

- Df for Model 2: The difference in residual degrees of freedom between Model 1 and Model 2, hence $2695 - 2694 = 1$
- Deviance for Model 2: The difference between the deviance of Model 1 and Model 2, hence $2996.8 - 2981.7 = 15.1$
- Resid. Df for Model 3: The residual degrees of freedom, which is the difference between the number of observations (2702) and the number of parameters in Model 3. The number of parameters is 1 (intercept)+ 6 (1 for each of the main effects) + 1 (year:patus) + 1 (patus:ncountry)= 9, hence Resid. Df for Model 3 is $2702 - 9 = 2693$
- Resid. Dev for Model 4: The deviance for the fit of Model 4, which can be found by subtracting the difference between the deviance of Model 3 and Model 4 (2.5899) from the deviance for the fot of Model 3 (2979.2), hence $2979.2 - 2.5899 = 2976.61$

(Continued on page 4.)

```
     > anova(fit1,fit2,fit3,fit4,test="LRT")
Analysis of Deviance Table

Model 1: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry
Model 2: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry + year:patus
Model 3: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry + year:patus +
    patus:ncountry
Model 4: opp ~ year + ncit + ustwin + patus + patgsgr + ncountry + year:patus +
    patus:ncountry + patgsgr:ncountry
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2695     2996.8
2      2694     2981.7  1     15.1 0.0001011 ***
3      2693     2979.2  1   2.5449 0.1106478
4      2692     2976.61 1   2.5899 0.1075493
```

c

The AIC for a model with p parameters and maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is defined as $-2\left(L(\hat{\boldsymbol{\beta}}) - p\right)$. Minimising this over a set of models identifies which of these models that has a distribution closest to the true distribution.

The likelihood ratio test can only be used for nested models, while comparing AIC values does not have this restriction. In this example the probit model achieves the smallest AIC value, which is indicates that it is the best choice of link function here. However, the differences are very small, in particular between probit and logit.

Problem 3

a

The log-likelihood function is

$$L(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \mid \mathbf{V} \mid -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Hence, using the hint and that both a scalar and $\mathbf{V}^{-1}$ are symmetric,

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2}\frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\frac{1}{2}\frac{\partial(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + \ \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\frac{1}{2}\frac{\partial(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - 2\mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= (\mathbf{y}^T\mathbf{V}^{-1}\mathbf{X})^T - \frac{1}{2}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^T\right)\boldsymbol{\beta} \\
&= \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

which is $= 0$ for

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

(Continued on page 5.)

b

Since both $\mathbf{Y}|\mathbf{u}$ and $\mathbf{u}$ are normally distributed, the joint distribution of $\mathbf{Y}$ and $\mathbf{u}$ is also normal. We have $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, $E[\mathbf{u}] = \mathbf{0}$, $\text{Var}[\mathbf{Y}] = \mathbf{V}$ and $\text{Var}[\mathbf{u}] = \boldsymbol{\Sigma}_\mathbf{u}$. Furthermore

$$\text{Cov}(\mathbf{Y}, \mathbf{u}) = \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \mathbf{u}) = \text{Cov}(\mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \mathbf{u}) = \mathbf{Z}\text{Cov}(\mathbf{u}, \mathbf{u}) = \mathbf{Z}\text{Var}(\mathbf{u}) = \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u}$$

which means that $\text{Cov}(\mathbf{Y}, \mathbf{u}) = (\mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u})^T = \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T$. Hence

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{u} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u} \\ \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T & \boldsymbol{\Sigma}_\mathbf{u} \end{pmatrix} \right]$$

Using the formula for conditional expectation for the multivariate normal distribution given in the appendix, we get

$$E[\boldsymbol{u}|\mathbf{Y} = \mathbf{y}] = \mathbf{0} + (\mathbf{Z}\boldsymbol{\Sigma}_\mathbf{u})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad (2)$$

Since $\mathbf{V}$ and $\boldsymbol{\Sigma}_\mathbf{u}$, are assumed known, a prediction of the random effects can be achieved by replacing $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}$ in (2)

$$\hat{\mathbf{u}} = \boldsymbol{\Sigma}_\mathbf{u}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

c

The marginal expected value of Y_{ij} is

$$\mu_{ij} = E[Y_{ij}] = E[E[Y_{ij}|\mathbf{u}_i]] = \int E[Y_{ij}|\mathbf{u}_i]f(\mathbf{u}_i)d\mathbf{u}_i = \int (\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i)f(\mathbf{u}_i)d\mathbf{u}_i$$
$$= \mathbf{x}_{ij}\boldsymbol{\beta}\int f(\mathbf{u}_i)d\mathbf{u}_i + \mathbf{z}_{ij}\int \mathbf{u}_i f(\mathbf{u}_i)d\mathbf{u}_i = \mathbf{x}_{ij}\boldsymbol{\beta}$$

since $\int f(\mathbf{u}_i)d\mathbf{u}_i = 1$ and $\int \mathbf{u}_i f(\mathbf{u}_i)d\mathbf{u}_i = E[\mathbf{u}_i] = \mathbf{0}$. Hence, as for the GLMM, the link function for the implied marginal model is also the identity link. This is not a general result for all lunk functions.

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100 / STK4100 — Introduction to generalized linear models.

Day of examination: Wednesday December 2nd 2020

Examination hours: 9.00 – 13.00.

This problem set consists of 5 pages.

Appendices: Formulas in STK3100 / STK4100

Permitted aids: All resources

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

Problem 1

a) We rewrite $\frac{\lambda^y}{y!}\exp(-\lambda) = \exp(y\log(\lambda) - \lambda - \log(y!))$ which gives $\theta = \log(\lambda)$, $\lambda = \exp(\theta) = b(\theta)$ and $c(y) = -\log(y!)$.

By general results $\mu = \mathrm{E}[Y] = b'(\theta) = \exp(\theta) = \lambda$ and $\mathrm{var}[Y] = b''(\theta) = \exp(\theta) = \lambda = \mu$.

b) For $y = 1, 2, 3, \ldots$ we have

$$\mathrm{P}(Y = y|Y > 0) = \frac{\mathrm{P}(Y = y)}{\mathrm{P}(Y > 0)} = \frac{\mathrm{P}(Y = y)}{1 - \mathrm{P}(Y = 0)} = \frac{\lambda^y \exp(-\lambda)/y!}{1 - \exp(-\lambda)}$$

Then rewrite $\frac{\lambda^y}{y!}\exp(-\lambda)/(1 - \exp(-\lambda)) = \exp(y\log(\lambda) - \lambda - \log(y!) - \log(1 - \exp(-\lambda))) = \exp(y\theta - b(\theta) + c(y))$ with $\theta = \log(\lambda)$ (as in a)), $b(\theta) = \lambda + \log(1 - \exp(-\lambda)) = \exp(\theta) - \log(1 - \exp(-\exp(\theta)))$ and $c(y) = -\log(y!)$ (also as in a)).

c) We have (when $f(y;\gamma)$ is a density, otherwise replace integral by sum) $\mathrm{P}(Y \in B) = \int_B f(y;\gamma)dy = \exp(-b_0(\gamma)) \int_B \exp(\gamma y - c_0(y))dy$. Thus $Y|Y \in B$ has a density

$$\begin{aligned} f_B(y;\theta) &= \frac{f(y;\gamma)}{\mathrm{P}(Y\in B)} = \frac{\exp(\gamma y - b_0(\gamma) + c_0(y))}{\exp(-b_0(\gamma)) \int_B \exp(\gamma y - c_0(y))dy} \\ &= \exp(\gamma y - \log(\textstyle\int_B \exp(\gamma y - c_0(y))dy) + c_0(y)) \end{aligned}$$

and so $\theta = \gamma, c(y) = c_0(y)$ and $b(\theta) = \log(\int_B \exp(\theta y - c(y))dy)$.

(Continued on page 2.)

Problem 2

a) The logistic regression model here is

$$\mathrm{P}(Y_i = 1|x_{i1}, x_{i2}) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2}}.$$

Thus we get the odds

$$\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2})} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2})$$

which lead to the *odds-ratio*

$$(\frac{\mathrm{P}(Y_i = 1|x_{i1} + 1, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1} + 1, x_{i2}))})/(\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}))}) = \exp(\beta_1)$$

as general interpretations of $\exp(\beta_1)$ as *odds-ratios* when changing x_{i1} by one unit keeping x_{i2} constant and similarly for $\exp(\beta_2)$

When all $\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})$ are small we have $1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}) \approx 1$ and so

$$(\frac{\mathrm{P}(Y_i = 1|x_{i1} + 1, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1} + 1, x_{i2}))})/(\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}))}) \approx \frac{\mathrm{P}(Y_i = 1|x_{i1} + 1, x_{i2})}{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})},$$

i.e. as a *relative risk*.

Inserting estimates $\hat{\beta}_j$ leads to estimated odds-ratios approximated by estimated relative risks. Here we get $\exp(\hat{\beta}_1) = \exp(2.20) = 9.02$ so as an approximation bad health increases the chance of frequent doctorial visits by a factor 9 (Actually this will be an exaggerated increase since $exp(2.20) = 9.02$ is a large value). Similarly $\exp(\hat{\beta}_2) = \exp(-0.338) = 0.71$, so after the health reform the proportions of women with frequent doctoral visits were approximately reduced by 30%.

b) Approximately the MLE $\hat{\beta}_j \sim \mathrm{N}(\beta_j, se_j^2)$ (by slight abuse of notation since se_j are statistics/random variables) and so

$$\begin{aligned} 0.95 &\approx \mathrm{P}(-1.96 < (\hat{\beta}_j - \beta_j)/se_j < 1.96) \\ &= \mathrm{P}(\hat{\beta}_j - 1.96se_j < \beta_j < \hat{\beta}_j + 1.96se_j) \\ &= \mathrm{P}(\exp(\hat{\beta}_j - 1.96se_j) < \exp(\beta_j) < \exp(\hat{\beta}_j + 1.96se_j)) \end{aligned}$$

Inserting the estimated regression coefficients and standard errors gives 95% confidence interval

$(6.51, 12.51)$ for $\exp(\beta_1)$

$(0.52, 0.98)$ for $\exp(\beta_2)$,

(Continued on page 3.)

none of which overlaps the value 1. Thus we can reject both null hypotheses H_{0j} at a 5 percent level. In particular the interval for $\exp(\beta_1)$ has a low end far from 1, indicating strong statistical significance. This is confirmed by the $|z_j| = |\hat{\beta}_j / se_j|$ being values larger than 2 and p-values less than 0.05 (in particular for $j = 1$).

c) Deviances are two times the difference between the log-likelihood with a specific model and the log-likelihood with a saturated model where fitted values $\tilde{y}_i$ are equal to observed values y_i.

Differences in deviances between two nested models, i.e. a smaller model is a special case of a larger, is chi-square distributed with degrees of freedom equal to the diiference in number of parameters between the model given that the smaller model is true.

The approximation to the χ^2 distribution stems for the differences in deviances being equal to twice the differences in log-likehoods betweens the models, as the saturated log-likelihood terms cancels out, and so is due to the properties of the likelihood ratio test.

A deviance table gives deviances, changes in deviances and changes in no. of parameters for a series of nested models. This gives the opportunity to test a series of models and evaluate which (often categorical) explanatory variables that are essential or non-essential for the outcome.

In the table below the questions marks have been replaced by the acutal numbers:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2226	1303.4	
badh	1	158.613	2225	1144.8	< 2.2e-16
reform	1	4.404	2224	1140.4	0.035849
educat	2	2.339	2222	1138.0	0.310536
inccat	2	8.641	2220	1129.4	0.013292
badh:reform	1	1.458	2219	1127.9	0.227285
badh:inccat	2	0.851	2217	1127.1	0.653313
educat:inccat	4	13.689	2213	1113.4	0.008357

Problem 3

a) We get

$$\mu = \mathrm{E}[Y] = \mathrm{E}[\exp(V)] = M_V(1) = \exp(\gamma * 1 + \frac{1}{2}\sigma^2 * 1^2) = \exp(\gamma + \frac{1}{2}\sigma^2).$$

Thus

$$\begin{aligned} \mathrm{var}[Y] &= \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 = M_V(2) - M_V(1)^2 = \exp(2\gamma + 2\sigma^2) - \exp(2\gamma + \sigma^2) \\ &= \mu^2(\exp(\sigma^2) - 1) = \phi\mu^2 \end{aligned}$$

(Continued on page 4.)

with $\phi = \exp(\sigma^2) - 1$.

b) Since $\mu_i = \exp(\alpha + \beta x_i + \frac{1}{2}\sigma^2) = \exp(\gamma_i + \frac{1}{2}\sigma^2)$ with $\gamma_i = \alpha + \beta x_i$ we get that

$$\mathrm{E}[V_i] = \mathrm{E}[\log(Y_i)] = \gamma_i = \alpha + \beta x_i$$

and so (α, β) can be estimated by least squares estimates $(\hat{\alpha}, \hat{\beta})$ of a simple linear regression on $V_i = \log(Y_i)$. Furthermore σ^2 can then be estimated as $\hat{\sigma}^2 = \sum_{i=1}^n (V_i - \hat{\alpha} - \hat{\beta} x_i)^2/(n-2)$ which then leads to estimate $\hat{\phi} = \exp(\hat{\sigma}^2) - 1$ of ϕ.

c) The score equations for generalized linear models can be written as

$$\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{Y_i - \mu_i}{\mathrm{var}(Y_i)} = \frac{1}{\phi} \sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{Y_i - \mu_i}{\nu^*(\mu_i)} = 0; \quad j = 1, \ldots, p$$

for models with $\mathrm{var}(Y_i) = \phi \nu^*(\mu_i)$.

These estimating equations can also be used as so-called quasi-likelihood equations under the weaker assumption that $Y_i; i = 1, \ldots, n$, are independent, but not necessarily from an exponential dispersion family, $g(\mu_i) = g(\mathrm{E}[Y_i]) = \beta' x_i$ for a link function $g()$ and variance specification $\mathrm{var}(Y_i) = \phi \nu^*(\mu_i)$. This leads to consistent and asymptotically normal estimates of β with a variance matrix as the inverse of the information matrix (- expected Jacobi) based on the quasi-score function.

In this particular case one obtains the estimates by specifying a gamma family with an identity link in the `glm`-command (since the variance function for the gamma family equals μ^2) or equivalently by specifying a quasi-family with identity link and μ^2 variance.

Problem 4

a) When $u_i \sim \mathrm{N}(0, \sigma^2)$ it has a density $f(u; \sigma_u^2) = \exp(-u^2/(2\sigma_u^2))/\sqrt{2\pi\sigma_u^2}$. Thus the marginal probability is by the rule of double expectation

$$\mathrm{P}(Y_{ij} = 1 | x_{ij}) = \mathrm{E}[\mathrm{P}(Y_{ij} = 1 | x_{ij}, u_i)] = \int \frac{\exp(\beta_0 + \beta_1 x_{ij} + u)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u)} f(u; \sigma_u^2) du$$

Similarily, for $j = 0, 1$ and $k = 0, 1$,

$$\begin{aligned}
\pi_i(j, k; \beta_0, \beta_2, \sigma_u^2) &= \mathrm{P}(Y_{i1} = j, Y_{i2} = k | x_{i1}, x_{i2}) \\
&= \mathrm{E}[\mathrm{P}(Y_{i1} = j, Y_{i2} = k | x_{i1}, x_{i2}, u_i)] \\
&= \mathrm{E}[\mathrm{P}(Y_{i1} = j | x_{i1}, x_{i2}, u_i) \mathrm{P}(Y_{i2} = k | x_{i1}, x_{i2}, u_i)] \\
&= \int \tfrac{\exp(j(\beta_0 + \beta_1 x_{i1} + u))}{1 + \exp(\beta_0 + \beta_1 x_{i1} + u)} \tfrac{\exp(k(\beta_0 + \beta_1 x_{i2} + u))}{1 + \exp(\beta_0 + \beta_1 x_{i2} + u)} f(u; \sigma_u^2) du
\end{aligned}$$

(Continued on page 5.)

which can not be written as $\mathrm{P}(Y_{i1}=1|x_{i1})\mathrm{P}(Y_{i2}=1|x_{i2})$. Thus Y_{i1} and Y_{i2} are marginally dependent.

The marginal likelihood can then be written

$$l(\beta_0,\beta_1,\sigma_u^2)=\prod_{i=1}^{n}\pi_i(Y_{i1},Y_{i2};\beta_0,\beta_1,\sigma_u^2).$$

b) If $Y_{i1}+Y_{i2}=2$ then necessarily both $Y_{i1}=1$ and $Y_{i2}=1$, thus $\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=2)=1$. Similarly, $Y_{i1}+Y_{i2}=0$ imply that both $Y_{i1}=0$ and $Y_{i2}=0$ and so also $\mathrm{P}(Y_{i1}=0|Y_{i1}+Y_{i2}=0)=1$. Thus no such pair (Y_{i1},Y_{i2}) will conditionallly on $Y_{i1}+Y_{i2}$ contain any information on β_1.

But when $Y_{i1}+Y_{i2}=1$ then either $Y_{i1}=1$ and $Y_{i2}=0$ or $Y_{i1}=0$ and $Y_{i2}=1$ and so, conditionally on u_i, x_{i1} and x_{2i},

$$\begin{aligned}\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1) &= \frac{\mathrm{P}_{(Y_{i1}=1,Y_{i2}=0)}}{\mathrm{P}_{(Y_{i1}=1,Y_{i2}=0)}+\mathrm{P}_{(Y_{i1}=0,Y_{i2}=1)}}\\ &= \frac{\exp(\beta_1(x_{i1}-x_{i2}))}{1+\exp(\beta_1(x_{i1}-x_{i2}))}\end{aligned}$$

since (same conditioning on u_i, x_{i1}, x_{i2})

$$\mathrm{P}(Y_{i1}=1,Y_{i2}=0)=\frac{\exp(\beta_0+\beta_1x_{i1}+u_i)}{1+\exp(\beta_0+\beta_1x_{i1}+u_i)}\frac{1}{1+\exp(\beta_0+\beta_1x_{i2}+u_i)}.$$

The expression for $\mathrm{P}(Y_{i1}=0,Y_{i2}=1)$ has the same denominator which then cancel out in $\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1)$. One is then left with

$$\begin{aligned}\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1) &= \frac{\exp(\beta_0+\beta_1x_{i1}+u_i)}{\exp(\beta_0+\beta_1x_{i1}+u_i)+\exp(\beta_0+\beta_1x_{i2}+u_i)}\\ &= \frac{\exp(\beta_1x_{i1})}{\exp(\beta_1x_{i2})+\exp(\beta_1x_{i2})}=\frac{\exp(\beta_1(x_{i1}-x_{i2}))}{1+\exp(\beta_1(x_{i1}-x_{i2}))}\end{aligned}$$

This means that it is possible to estimate β_1 by running a logistic regression

- with outcome Y_{i1}
- for pairs with $Y_{i1}+Y_{i2}=1$
- with explanatory variables $x_{i1}-x_{i2}$
- and no intercept

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100/STK4100 — Introduction to generalized linear models.

Day of examination: Wednesday December 18th 2019

Examination hours: 9.00 – 13.00.

This problem set consists of 0 pages.

Appendices: Formulas for STK3100 and STK4100

Permitted aids: Approved calculator

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

Problem 1

a) The odds is defined as

$$\text{Odds}(x_i) = \frac{\pi_i}{1-\pi_i} = \frac{\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}}{\frac{1}{1+\exp(\alpha+\beta x_i)}} = \exp(\alpha+\beta x_i)$$

Then we can write an odds-ratio between observations with explanatory variables $x_i' = x_i + 1$ and x_i as

$$\text{OR} = \frac{\text{Odds}(x_i+1)}{\text{Odds}(x_i)} = \frac{\exp(\alpha+\beta(x_i+1))}{\exp(\alpha+\beta x_i)} = \exp(\beta)$$

When both π_i and $\pi_{i'}$ are small then $1-\pi_i \approx 1$ and $1-\pi_{i'} \approx 1$. Thus $\text{Odds}(x_i) \approx \pi_i$ and $\exp(\beta) = \text{OR} \approx \pi_{i'}/\pi_i = \text{RR}$, i.e. a relative risk

b) We have

$$\text{P}(Y_i = 1|Z_i = 1) = \frac{\text{P}(Y_i = 1 \cap Z_i = 1)}{\text{P}(Z_i = 1)}$$

where $\text{P}(Y_i = 1 \cap Z_i = 1) = \text{P}(Z_i = 1|Y_i = 1)\text{P}(Y_i = 1) = \rho_1\pi_i$.

Similarily $\text{P}(Y_i = 0 \cap Z_i = 1) = \text{P}(Z_i = 1|Y_i = 0)\text{P}(Y_i = 0) = \rho_0(1-\pi_i)$.

Thus $\text{P}(Z_i = 1) = \rho_1\pi_i + \rho_0(1-\pi_i)$ and

$$\text{P}(Y_i = 1|Z_i = 1) = \frac{\rho_1\pi_i}{\rho_1\pi_i + \rho_0(1-\pi_i)} = \frac{\rho_1\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}}{\rho_1\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)} + \rho_0\frac{1}{1+\exp(\alpha+\beta x_i)}}$$

(Continued on page 2.)

which simplifies to

$$\mathrm{P}(Y_i = 1|Z_i = 1) = \frac{\rho_1 \exp(\alpha + \beta x_i)}{\rho_1 \exp(\alpha + \beta x_i) + \rho_0} = \frac{\exp(\alpha^* + \beta x_i)}{1 + \exp(\alpha^* + \beta x_i)}$$

with $\alpha^* = \alpha + \log(\rho_1/\rho_0)$.

The implication of the result is that the same odds-ratio $\exp(\beta)$ can be estimated both on cohort (population) and on case-control data.

Problem 2

a) The gamma density can be rewritten as

$$\begin{aligned} f(y;\mu,k) &= \exp(-\tfrac{k}{\mu}y - k\log(\mu) + k\log(k) + (k-1)\log(y) - \log(\Gamma(k))) \\ &= \exp(\tfrac{(-1/\mu)y - \log(\mu)}{1/k} - k\log(k) + (k-1)\log(y) - \log(\Gamma(k))) \\ &= \exp((\theta y - b(\theta))/\phi + c(y,\phi)) \end{aligned}$$

with $\theta = -1/\mu, b(\theta) = \log(\mu) = -\log(1/\mu) = -\log(-\theta), \phi = 1/k$ and $c(y,\phi) = -\log(1/\phi)/\phi + (1/\phi - 1)\log(y) - \log(\Gamma(1/\phi))$.

We have $\mathrm{E}[Y] = b'(\theta) = -\frac{1}{-\theta}(-1) = -\frac{1}{\theta} = \mu$ and $\mathrm{var}[Y] = \phi b''(\theta) = \phi\frac{1}{\theta^2} = \phi\mu^2$.

b) A GLM consists of three components

(i) Independent Y_i from a distribution $\exp((\theta_i y - b(\theta_i))/\phi + c(y,\phi))$ with $\mu_i = b'(\theta_i)$

(ii) A linear predictor $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$

(iii) A link function $g(\mu_i) = \eta_i$.

We saw in a) that (i) was satisfied, so a GLM for gamma distributed Y_i thus requires specification of (ii) the linear predictor η_i and (iii) the link function $g()$.

The log-likelihood can be written as $L(\beta) = \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i))/\phi + c(y_i,\phi)$. By the chain rule the derivatives of $L(\beta)$ then becomes

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \eta_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial(y_i\theta_i - b(\theta_i))}{\partial \theta_i}\frac{1}{\phi} = \sum_{i=1}^{n} x_{ij}\frac{\partial \mu_i}{\partial \eta_i}\frac{(y_i - \mu_i)}{\phi}\frac{\partial \theta_i}{\partial \mu_i}$$

and since $\frac{\partial \theta_i}{\partial \mu_i} = 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/b''(\theta_i)$ and $\mathrm{var}[Y_i] = \phi b''(\theta) = \phi\mu_i^2$ for the gamma family we obtain

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi\mu_i^2} x_{ij}\frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for } j = 1,\ldots,p$$

(Continued on page 3.)

c) Since $\mathrm{E}[Y_i] - \mu_i = 0$ as long as the μ_i are correctly specified the score equations are still unbiased. The construction is called the (score equations) for quasi-likelihood which only require the structure of expectation and variance to be correctly specified to give asymptotically normal estimators with expected information matrix equal to the covariance matrix of the scores.

The dispersion parameter ϕ can be estimated using the Pearson X^2 with the particular variance structure, thus we can use

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$$

as a consistent estimator of ϕ when $\mathrm{var}[Y_i] = \phi\mu_i^2$.

d) We see that the price depends significantly on x_1 size, x_4 rent, x_5 distance to the east (x) and also to whether the appartement has a balcony (x_3), but that the number of rooms (x_2) does not a significantly effect (when adjusting for size). Increase of x_1 and x_3 increases the price, whereas increase in x_4 and x_5 decreases the price.

The natural estimate for $\mu = \eta = \beta_0 + \sum_{j=1}^{5}\beta_j x_j = \beta^t\mathbf{x}$ equals $\hat{\mu} = \hat{\beta}_0 + \sum_{j=1}^{5}\hat{\beta}_j x_j = 526.64 + 18.4*70 + 25.77*2 - 0.12745*1000 - 93.29*2 =$ 1552 x 1000 NOK.

With $\hat{\mu} = \hat{\beta}'\mathbf{x}$ we can estimate $\mathrm{var}(\hat{\mu})$ by $\mathbf{x}^t\hat{\Sigma}\mathbf{x}$ where $\hat{\Sigma}$ is the estimated covariance matrix for $\hat{\beta}$.

e) The linear regression model used here can be expressed as

$$Y_i = \beta_0 + \sum_{j=1}^{5}\beta_j x_{ij} + \varepsilon_i$$

where the $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ and independent.

Roughly the $\hat{\beta}_j$ and p-values correspond well between the gamma-regression and the usual linear regression model. The residual plots of (deviance) residuals vs. fitted values reveal no clear non-linearities for either model. However, it seems that the residuals for the linear regression model tend to increase as the fitted values increases. It appears that for the gamma model the deviance residuals does not display such a tendency. It may thus be that the gamma model with variance structure $\phi\mu^2$ captures heteroscedasticity better than the constant variance in the linear regression.

(Although the homoscedastic model did not affect estimates or p-values much it will certainly be an important issue if we want prediction intervals for predicted prizes with a given new vector of explanatory variables $\mathbf{x}$. But this perspective was not discussed seriously in STK3100/4100 this semester).

(Continued on page 4.)

Problem 3

a) We have $\mathrm{E}[Y_{ij}] = \mathrm{E}[\mathrm{E}(Y_{ij}|u_i)] = \mathrm{E}[\exp(\beta_0 + \beta_1 x_{ij} + u_i)] = \exp(\beta_0 + \beta_1 x_{ij})\mathrm{E}[e^{u_i}]$ since $\beta_0 + \beta_1 x_{ij}$ is a constant, not a random variable.

Also, $\mathrm{E}[e^{u_i}] = M(1) = \exp(\sigma_u^2/2)$, thus marginally $\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$ and only the intercept β_0 is changed relative to the generalized linear mixed effects model.

b) We have that $\mathrm{var}[Y_{ij}] = \mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|u_i)]$. Here $\mathrm{var}(Y_{ij}|u_i)] = \exp(\beta_0 + \beta_1 x_{ij} + u_i) = \mathrm{E}(Y_{ij}|u_i)]$ and so $\mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$.

Furthermore, $\mathrm{var}[\mathrm{E}(Y_{ij}|u_i)] = \mathrm{var}(\exp(\beta_0 + \beta_1 x_{ij} + u_i)) = \exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{var}(\exp(u_i)) = \exp(2(\beta_0 + \beta_1 x_{ij})\ [M(2) - M(1)^2]$ and $M(2) - M(1)^2 = \exp(2\sigma_u^2) - \exp(\sigma_u^2) = \exp(\sigma_u^2)(\exp(\sigma_u^2) - 1)$. Thus

$$\mathrm{var}[Y_{ij}] = \exp(\beta_0+\beta_1 x_{ij}+\sigma_u^2/2)+\exp(2(\beta_0+\beta_1 x_{ij}))\exp(\sigma_u^2)(\exp(\sigma_u^2)-1).$$

c) Also when Y_{ij} given u_i is gamma distributed we get $\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$ by the same derivation as in a). The result that only the intercept changes from the mixed to the marginal model depends only on the log-link structure.

For the marginal variances of the Y_{ij} we again have $\mathrm{var}[Y_{ij}] = \mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|u_i)]$ and

$$\begin{aligned}\mathrm{var}[\mathrm{E}(Y_{ij}|u_i)] &= \mathrm{var}(\exp(\beta_0 + \beta_1 x_{ij} + u_i)) \\ &= \exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{var}(\exp(u_i)) \\ &= \exp(2(\beta_0 + \beta_1 x_{ij}))\exp(\sigma_u^2)(\exp(\sigma_u^2) - 1)\end{aligned}$$

as in question b).

Furthermore, $\mathrm{var}(Y_{ij}|u_i) = \phi\mu_{ij}^2 = \phi\exp(2(\beta_0 + \beta_1 x_{ij} + u_i))$, so

$$\begin{aligned}\mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] &= \mathrm{E}[\phi\exp(2(\beta_0 + \beta_1 x_{ij} + u_i))] \\ &= \phi\exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{E}[\exp(2u_i)] \\ &= \phi\exp(2(\beta_0 + \beta_1 x_{ij}))M(2) = \phi\exp(2(\beta_0 + \beta_1 x_{ij}) + 2\sigma_u^2)\end{aligned}$$

As in b) the answer is then obtained adding these two terms.

END

(Continued on page 5.)

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to generalized linear models. SOLUTIONS TO PROBLEMS
Day of examination:	Friday 14 December 2018.
Examination hours:	09.00 – 13.00.
This problem set consists of 9 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

We have $V \sim \text{bin}(n, \pi)$ and $Y = V/n$. The pmf of Y is given by

$$P(Y = y) = \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \tag{1}$$

for $y = 0, \frac{1}{n}, \frac{2}{n} \ldots . \frac{n-1}{n}, 1$.

a) We may rewrite (1) as

$$\begin{aligned} P(Y = y) &= \exp\left\{ ny \log(\pi) + (n - ny)\log(1-\pi) + \log\binom{n}{ny} \right\} \\ &= \exp\left\{ \frac{y \log\left(\frac{\pi}{1-\pi}\right) - [-\log(1-\pi)]}{1/n} + \log\binom{n}{ny} \right\}. \end{aligned}$$

This is of the form (2) in the exam problems with $\theta = \log\left(\frac{\pi}{1-\pi}\right)$. Hence $\pi = e^\theta/(1+e^\theta)$, and we have that $b(\theta) = -\log(1-\pi) = \log(1+e^\theta)$, $a(\phi) = 1/n$ and $c(y, \phi) = \log\binom{n}{ny}$.

b) We have that

$$\mu = E(Y) = b'(\theta) = \frac{d}{d\theta}\log(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi,$$

and

$$\begin{aligned} \text{Var}(Y) &= b''(\theta) \cdot a(\phi) = \frac{d}{d\theta}\left(\frac{e^\theta}{1+e^\theta}\right) \cdot \frac{1}{n} = \frac{e^\theta}{(1+e^\theta)^2} \cdot \frac{1}{n} \\ &= \frac{1}{n} \cdot \frac{e^\theta}{1+e^\theta} \cdot \frac{1}{1+e^\theta} = \frac{1}{n}\pi(1-\pi). \end{aligned}$$

(Continued on page 2.)

We then assume that $V_1, V_2, \ldots, V_N$ are independent with $V_i \sim \text{bin}(n_i, \pi_i)$, and let $Y_i = V_i/n_i$ for $i = 1, 2, \ldots, N$. We consider a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_N$ with canonical link function $\text{logit}(\pi_i) = \log\{\pi_i/(1-\pi_i)\} = \eta_i$ and linear predictor $\eta_i = \beta_0 + \beta_1 x_i$. Here $x_1, \ldots, x_N$ are known covariate values.

c) Let $y_1, y_2, \ldots, y_N$ be the observed values of $Y_1, Y_2, \ldots, Y_N$. Then the likelihood is given by

$$\ell(\beta_0, \beta_1) = \prod_{i=1}^N \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1-\pi_i)^{n_i - n_i y_i} = C \prod_{i=1}^N \left(\frac{\pi_i}{1-\pi_i}\right)^{n_i y_i} (1-\pi_i)^{n_i},$$

where $C = \prod_{i=1}^N \binom{n_i}{n_i y_i}$. Now $\pi_i = e^{\eta_i}/(1+e^{\eta_i})$, so the log-likelihood becomes

$$\begin{aligned} L(\beta_0, \beta_1) &= \log C + \sum_{i=1}^N \left\{ n_i y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) \right\} \\ &= \log C + \sum_{i=1}^N \left\{ n_i y_i \eta_i + n_i \log\left(\frac{1}{1+e^{\eta_i}}\right) \right\} \\ &= \log C + \sum_{i=1}^N \left\{ n_i y_i (\beta_0 + \beta_1 x_i) - n_i \log(1 + e^{\beta_0 + \beta_1 x_i}) \right\} \end{aligned}$$

We differentiate the log-likelihood function, and find

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^N \left(n_i y_i - n_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = \sum_{i=1}^N n_i (y_i - \pi_i),$$

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^N \left(n_i y_i x_i - n_i \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = \sum_{i=1}^N n_i x_i (y_i - \pi_i).$$

We obtain the maximum likelihood estimates by solving the equation we obtain by setting the partial derivatives equal to zero. Expressed in terms of the random Y_i's we therefore have that the maximum likelihood estimators are the solutions of the equations

$$\sum_{i=1}^N n_i (Y_i - \pi_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^N n_i x_i (Y_i - \pi_i) = 0.$$

d) We first note that

$$\begin{aligned} \frac{\partial \pi_i}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = \frac{e^{\beta_0 + \beta_1 x_i}}{(1 + e^{\beta_0 + \beta_1 x_i})^2} \\ &= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \cdot \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} = \pi_i (1 - \pi_i), \end{aligned}$$

(Continued on page 3.)

and

$$\begin{aligned}\frac{\partial \pi_i}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) = \frac{x_i e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \\ &= x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \cdot \frac{1}{1+e^{\beta_0+\beta_1 x_i}} = x_i \pi_i (1-\pi_i).\end{aligned}$$

By differentiating the log-likelihood function one more time, we then find that

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_0^2} = \frac{\partial}{\partial \beta_0}\sum_{i=1}^N n_i(y_i-\pi_i) = -\sum_{i=1}^N n_i \frac{\partial \pi_i}{\partial \beta_0} = -\sum_{i=1}^N n_i \pi_i (1-\pi_i),$$

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_1^2} = \frac{\partial}{\partial \beta_1}\sum_{i=1}^N n_i x_i(y_i-\pi_i) = -\sum_{i=1}^N n_i x_i \frac{\partial \pi_i}{\partial \beta_1} = -\sum_{i=1}^N n_i x_i^2 \pi_i (1-\pi_i),$$

and

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_0} = \frac{\partial}{\partial \beta_1}\sum_{i=1}^N n_i(y_i-\pi_i) = -\sum_{i=1}^N n_i \frac{\partial \pi_i}{\partial \beta_1} = -\sum_{i=1}^N n_i x_i \pi_i (1-\pi_i).$$

The second order patrial derivatives do not depend on the y_i's. Therefore the observed and the expected information matrices coincided, and we have that

$$\boldsymbol{\mathcal{J}} = \left\{-E\left(\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_h \partial \beta_j}\right)\right\}_{h,j=0,1} = \begin{pmatrix} \sum_{i=1}^N n_i \pi_i (1-\pi_i) & \sum_{i=1}^N n_i x_i \pi_i (1-\pi_i) \\ \sum_{i=1}^N n_i x_i \pi_i (1-\pi_i) & \sum_{i=1}^N n_i x_i^2 \pi_i (1-\pi_i) \end{pmatrix}.$$

Problem 2

a) The analysis reported in question a is based on a logistic regression model. To describe the model, we let $Y_i = 1$ if passenger i survived the disaster, $Y_i = 0$ if passenger i died, and we let $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$ denote its covariates (including $x_{i0} = 1$ for the intercept) defined as follows:

$x_{i1} = 1$ if passenger i is female; $x_{i1} = 0$ if passenger i is male,

$x_{i2} = a_i - 30$, where a_i is the age (in years) of passenger i,

$x_{i3} = 1$ if passenger i travelled on second class; $x_{i3} = 0$ otherwise,

$x_{i4} = 1$ if passenger i travelled on third class; $x_{i4} = 0$ otherwise.

The model assumes that the Y_i's are independent and that $\pi_i = P(Y_i = 1)$ is given as

$$\pi_i = \frac{\exp\left(\sum_{j=0}^4 \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^4 \beta_j x_{ij}\right)}.$$

(Continued on page 4.)

To give an interpretation of the estimated intercept, we consider a 30 years old male passenger who travelled on first class. Such a passenger has $x_{ij} = 0$ for $j = 1, 2, 3, 4$, so (according to the given model) his probability of surviving the disaster is $\pi_i = e^{\beta_0}/(1 + e^{\beta_0})$, and this is estimated by

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0}}{1 + e^{\widehat{\beta}_0}} = \frac{e^{-0.007567}}{1 + e^{-0.007567}} = 0.498.$$

This gives an interpretation of the estimated intercept.

To interpret the estimate for (centered) age, we consider two passengers, k and i, of the same sex who travelled on the same class, but where passenger k is one year older than passenger i. Then the odds ratio for the two passengers is

$$\mathrm{OR}(k, i) = \frac{\pi_k/(1 - \pi_k)}{\pi_i/(1 - \pi_i)} = \frac{\exp\left(\sum_{j=0}^{4} \beta_j x_{kj}\right)}{\exp\left(\sum_{j=0}^{4} \beta_j x_{ij}\right)} = e^{\beta_2},$$

and this is estimated by

$$\widehat{\mathrm{OR}}(k, i) = e^{\widehat{\beta}_2} = e^{-0.034393} = 0.966.$$

Thus the odds of surviving is reduced by 3.4% when the age is increased by one year.

b) The model in question a has (residual) deviance 982.45 with 1041 degrees of freedom, while the (residual) deviance of the model in question b is 931.99 with 1039 degrees of freedom. The difference in (residual) deviance between the two models is $982.45 - 931.99 = 50.46$, while the difference in the degrees of freedom is $1041 - 1039 = 2$. The difference between the deviances equals minus two times the likelihood ratio of the model in a to the model in b. So it is approximately chi-squared distributed with 2 degrees of freedom if the model in a holds. Comparing the difference 50.46 between the deviances with a chi-squared distribution with 2 degrees of freedom, we see that the model in b gives an improved fit compared to the model in a.

For the model in question b there is interaction between sex and passenger class. So in addition to the covariates in question a, we for this model also have the covariates:

$x_{i5} = 1$ if passenger i is female who travelled on second class; $x_{i5} = 0$ otherwise,

$x_{i6} = 1$ if passenger i is female who travelled on third class; $x_{i6} = 0$ otherwise.

The expression for $\pi_i = P(Y_i = 1)$ now becomes

$$\pi_i = \frac{\exp\left(\sum_{j=0}^{6} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{6} \beta_j x_{ij}\right)}.$$

In order to describe the effects of sex and passenger class, we consider a passenger that is 30 years old (which makes $x_{i2} = 0$).

(Continued on page 5.)

For a 30 years old male the probability of surviving the disaster is estimated to be:

- For a male at first class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0}}{1+e^{\widehat{\beta}_0}} = \frac{e^{-0.234083}}{1+e^{-0.234083}} = 0.442.$$

- For a male at second class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_3}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_3}} = \frac{e^{-0.234083-1.600280}}{1+e^{-0.234083-1.600280}} = 0.138.$$

- For a male at third class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_4}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_4}} = \frac{e^{-0.234083-1.576159}}{1+e^{-0.234083-1.576159}} = 0.141$$

For a 30 years old female the probability of surviving the disaster is estimated to be:

- For a female at first class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1}} = \frac{e^{-0.234083+3.886388}}{1+e^{-0.234083+3.886388}} = 0.975.$$

- For a female at second class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_3+\widehat{\beta}_5}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_3+\widehat{\beta}_5}} = \frac{e^{-0.234083+3.886388-1.600280+0.070407}}{1+e^{-0.234083+3.886388-1.600280+0.070407}} = 0.893.$$

- For a female at third class:

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_4+\widehat{\beta}_6}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_4+\widehat{\beta}_6}} = \frac{e^{-0.234083+3.886388-1.576159-2.488805}}{1+e^{-0.234083+3.886388-1.576159-2.488805}} = 0.398.$$

We see that for all three classes, the probability that a female will survive is higher than for a male. For males the probability for surviving is highest for a first class passenger (44.2%) and much lower for a second or third class passenger (13.8% and 14.1 %, respectively). For females the probability of surviving is high both for first and second class (97.5% and 89.3%, respectively) and much lower for a third class passenger (39.8%).

c) The analysis of deviance table with numbers filled in for the question marks is given below. The numbers that have been filled in are underlined.

```
Model 1: Survived~Sex+Cage+Class+Sex:Class
Model 2: Survived~Sex+Cage+Class+Sex:Class+Cage:Class
Model 3: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage
Model 4: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage+Sex:Cage:Class
     Resid. Df    Resid. Dev    Df    Deviance    Pr(>Chi)
1         1039        931.99
2         1037        922.17     2        9.82     0.00739
3         1036        917.84     1      4.3308     0.03743
4         1034        915.97     2      1.8661     0.39335
```

(Continued on page 6.)

In order to describe how we have arrived at the underlined numbers, we first describe the content of columns two to five in the analysis of deviance table:

- `Resid.Df` is the residual degrees of freedom and equals $N - p$, where N is the number of observations (here $N = 1046$) and p is the number of parameters in the model.
- `Resid.Dev` is the deviance for the fitted model. (Note that what is denoted 'deviance' in the text book, is denoted 'residual deviance' by `R`.)
- `Df` is the difference in residual degrees of freedom for the model on the line above and the actual model. Note that this is the same as the difference in the number of parameters of the actual model and the model on the line above.
- `Deviance` is the difference in deviance for the model on the line above and the actual model.

Using this, we find the underlined numbers as follows:

- `Resid.Df` for model 3: Model 3 has 1 parameter for the intercept, 1 parameter for the main effect of `Sex`, 1 parameter for main effect of `Cage`, 2 parameters for the main effect of `Class`, 2 parameters for the interaction `Sex:Class`, 2 parameters for the interaction `Cage:Class`, and 1 parameter for the interaction `Sex:Cage`. In total this gives 10 parameters, so `Resid.Df` equals $1046 - 10 = 1036$. [Alternatively (and easier) we may use that `Resid.Df` for model 2 is 1037 and the difference in residual degrees of freedom between models 2 and 3 is 1, so `Resid.Df` for model 3 is $1037 - 1 = 1036$.]
- `Resid.Dev` for model 4: From the output, the (residual) deviance of model 3 is 917.84 and the difference in (residual) deviance between models 3 and 4 is 1.8661. Hence the (residual) deviance for model 4 is $917.84 - 1.87 = 915.97$.
- `Df` for model 2: The difference in residual degrees of freedom for models 1 and 2 is $1039 - 1037 = 2$.
- `Deviance` for model 2: The difference in (residual) deviance between models 1 and 2 is $931.99 - 922.17 = 9.82$.

To compare the models, we may look at the P-values in the last column of the analysis of deviance table. (The P-value for a line in the table, is the P-value for a likelihood ratio test that tests the null hypothesis that the model on the previous line of the table holds, assuming that the model on the given line holds.) From the P-values, we see that model 2 gives a significantly better fit than model 1, and that model 3 gives a significantly better fit than model 2. But model 4 does not significantly improve the fit compared to model 3. Therefore we prefer model 3.

(Continued on page 7.)

Problem 3

a) The projection matrix is given by $\boldsymbol{P}_1 = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. To evaluate this, we first note that

$$\begin{aligned}
\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} &= \begin{bmatrix} \mathbf{1}_n^{\mathrm{T}} \\ (\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}} \end{bmatrix} [\mathbf{1}_n\,,\, \boldsymbol{x} - \bar{x}\mathbf{1}_n] \\
&= \begin{bmatrix} \mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n & \mathbf{1}_n^{\mathrm{T}}(\boldsymbol{x} - \bar{x}\mathbf{1}_n) \\ (\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}\mathbf{1}_n & (\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}(\boldsymbol{x} - \bar{x}\mathbf{1}_n) \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_{i=1}^n (x_i - \bar{x}) \\ \sum_{i=1}^n (x_i - \bar{x}) & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix} \\
&= \begin{bmatrix} n & 0 \\ 0 & M \end{bmatrix},
\end{aligned}$$

and therefore

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = \begin{bmatrix} n^{-1} & 0 \\ 0 & M^{-1} \end{bmatrix}.$$

We then find that the projection matrix may be given as

$$\begin{aligned}
\boldsymbol{P}_1 &= \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} \\
&= [\mathbf{1}_n\,,\, \boldsymbol{x} - \bar{x}\mathbf{1}_n] \begin{bmatrix} n^{-1} & 0 \\ 0 & M^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}_n^{\mathrm{T}} \\ (\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}} \end{bmatrix} \\
&= \left[n^{-1}\mathbf{1}_n\,,\, M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)\right] \begin{bmatrix} \mathbf{1}_n^{\mathrm{T}} \\ (\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}} \end{bmatrix} \\
&= n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}
\end{aligned}$$

b) From the result in a, we find the vector of fitted values may be given as

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \boldsymbol{P}_1\boldsymbol{Y} \\
&= \left[n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}\right]\boldsymbol{Y} \\
&= n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\boldsymbol{Y} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}\boldsymbol{Y} \\
&= \mathbf{1}_n n^{-1}\sum_{i=1}^n Y_i + (\boldsymbol{x} - \bar{x}\mathbf{1}_n)M^{-1}\sum_{i=1}^n Y_i(x_i - \bar{x}) \\
&= \overline{Y}\mathbf{1}_n + \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n).
\end{aligned}$$

(Continued on page 8.)

The projection matrix for the null model $\mu_i = E(Y_i) = \beta_0$ is $\boldsymbol{P}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$. We consider the orthogonal decomposition

$$\boldsymbol{Y} = \boldsymbol{P}_0\boldsymbol{Y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}$$

with corresponding sum of squares decomposition

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}.$$

c) Using the result in question a, we have that

$$\begin{aligned}
\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} &= \boldsymbol{Y}^{\mathrm{T}}[M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}]\boldsymbol{Y} \\
&= M^{-1}[\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)]\,[(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}}\boldsymbol{Y}] \\
&= M^{-1}\left[\sum_{i=1}^{n} Y_i(x_i - \bar{x})\right]^2 \\
&= M\widehat{\beta}_1^2,
\end{aligned}$$

which shows the first result. For the second result, we note that since $\boldsymbol{I} - \boldsymbol{P}_1$ is a projection matrix, it is idempotent and symmetric. Hence we have that

$$\boldsymbol{I} - \boldsymbol{P}_1 = (\boldsymbol{I} - \boldsymbol{P}_1)(\boldsymbol{I} - \boldsymbol{P}_1) = (\boldsymbol{I} - \boldsymbol{P}_1)^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1).$$

Then by the result in question b, we have that

$$\begin{aligned}
\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} &= \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} \\
&= [(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}]^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} \\
&= (\boldsymbol{Y} - \boldsymbol{P}_1\boldsymbol{Y})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{P}_1\boldsymbol{Y}) \\
&= [\boldsymbol{Y} - \overline{Y}\mathbf{1}_n - \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n)]^{\mathrm{T}}[\boldsymbol{Y} - \overline{Y}\mathbf{1}_n - \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n)] \\
&= \sum_{i=1}^{n}\left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2,
\end{aligned}$$

which shows the second result.

d) By Cochran's theorem we have that the terms on the right-hand side of (4) in the exam problems are independent, and when divided by σ_2 they are (non-central) chi-squared distributed. In particular, it follows that

$$M\widehat{\beta}_1^2/\sigma^2 = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y}/\sigma^2$$

and

$$\sum_{i=1}^{n}\left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2/\sigma^2 = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2$$

(Continued on page 9.)

are independent and (non-cental) chi-squared distributed.

Further, for (5) in the exam problems the degrees of freedom is given by (using that the rank of a projection matrix equals is trace)

$$\begin{aligned} \mathrm{df}_1 &= \mathrm{rank}(\boldsymbol{P}_1 - \boldsymbol{P}_0) = \mathrm{trace}(\boldsymbol{P}_1 - \boldsymbol{P}_0) \\ &= \mathrm{trace}(\boldsymbol{P}_1) - \mathrm{trace}(\boldsymbol{P}_0) = \mathrm{rank}(\boldsymbol{P}_1) - \mathrm{rank}(\boldsymbol{P}_0) \\ &= 2 - 1 = 1, \end{aligned}$$

while the degrees of freedom for (6) equals

$$\begin{aligned} \mathrm{df}_2 &= \mathrm{rank}(\boldsymbol{I} - \boldsymbol{P}_1) = \mathrm{trace}(\boldsymbol{I} - \boldsymbol{P}_1) \\ &= \mathrm{trace}(\boldsymbol{I}) - \mathrm{trace}(\boldsymbol{P}_1) = \mathrm{trace}(\boldsymbol{I}) - \mathrm{rank}(\boldsymbol{P}_1) \\ &= n - 2. \end{aligned}$$

The mean vector $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ is in the model space $C(\boldsymbol{X})$, so $\boldsymbol{P}_1\boldsymbol{\mu} = \boldsymbol{\mu}$. Therefore, by Cochran's theorem, the non-centrality parameter of (6) is

$$\lambda_2 = \frac{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{\mu}}{\sigma^2} = \frac{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{P}_1\boldsymbol{\mu})}{\sigma^2} = \frac{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu})}{\sigma^2} = 0$$

e) For testing the the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_A : \beta_1 \neq 0$, we may use the test statistic

$$F = \frac{\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y}/1}{\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}/(n-2)} = \frac{(n-2)M\widehat{\beta}_1^2}{\sum_{i=1}^n \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2}.$$

By the results of question d (including the one given in parenthesis), this has a non-central F-distribution with 1 and $n-2$ degrees of freedom and non-centrality parameter $\lambda_1 = M\beta_1^2/\sigma^2$. In particular, under H_0, we have $\lambda_1 = 0$ and then F has a central F-distribution with 1 and $n-2$ degrees of freedom.

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

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

a) We may rewrite (1) as

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

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

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

- $Y_1, Y_2, \ldots, Y_n$ 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 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, 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

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

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

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

(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

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

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 $\mathrm{var}(Y^*) = a(\phi)b''(\theta)$. 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

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

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

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

and

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

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 $\text{var}(Y_i)$ 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

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

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

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

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

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

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 $\Phi(z) = P(Z \leq z)$, we may write

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

(Continued on page 8.)

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

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

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

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

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.

Solution proposal finals STK3100/4100-f16

Problem 1

a) The density can be written

$$\begin{aligned} f(y;\mu,\nu) &= \frac{y^{-1}}{\Gamma(\nu)}(\frac{y\nu}{\mu})^{\nu}\exp(-y\nu/\mu), y>0. \\ &= \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}\exp(-y\nu/\mu-\nu\log(\mu)) \\ &= \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}\exp(\frac{y(-\frac{1}{\mu})-\log(\mu)}{\frac{1}{\nu}}). \end{aligned}$$

from which we see that $\phi = 1/\nu$, $c(y;\phi) = \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}$, $\theta = -1/\mu$ and $a(\theta) = \log(\mu) = \log(-1/\theta) = -\log(-\theta)$.

Since $a'(\theta) = -\frac{1}{\theta} = \mu$, $E(y) = \mu$.

b) The canonical link is obtained from $\theta = \eta$ where η is the predictor. The link is given by $\eta = g(\mu)$ so $-1/\theta = \mu = g^{-1}(\eta) = g^{-1}(\theta)$. Hence $g(-1/\theta) = \theta$ or $g(\theta) = -1/\theta$, i.e. the inverse. The problem with this link is that since $\mu > 0$, $\theta < 0$, the linear predictor will also be negative and more importantly not having the entire real line as range. This is not a good property, so the canonical link is not much used for gamma distributed response. Instead the log -link is much used.

Problem 2

a) The number of persons in each combination of the covariates is large. One can then think of the number of accidents as the sum of a large number of Bernoulli trials where the number of trials is large, and the success parameter, in this case the probability of being killed in a traffic accident, is small. The sum of the successes of Bernoulli trials has a Binomial distribution. For small success probabilities and large number of trials the probabilities in the Binomial distribution are close to the probabilities in a Poisson distribution. Hence it is reasonable to consider the responses as Poisson distributed in this case.

The number of groups is $2 \times 8 = 16$ and the number of parameters is $1+(2-1)+(8-1) = 9$ which means that the deviance is approximately χ^2-distributed with 16-9=7 degrees of freedom, cf. de Jong and Heller page 72. Then the probability for a value larger than the observed deviance is 0.19, so the fit is satisfactory.

b) The expected number of deaths in each group will depend on the size of the population. If $n_{ij}, i = 1,2,\ j = 1,\ldots,8$, are the population sizes, the expected number of deaths will be $n_{ij}f(gender_i, age_j)$.

Using the log link where $\eta = \log(\mu)$, or $\mu = \exp(\eta)$, the expected number of deaths in group ij will have the form $(n_{ij}/sc)\exp(\eta) = \exp(\log(n_{ij}/sc) + \eta)$. Remark that $exp(\eta)$ will have the interpretation as the rate pr sc units. The coefficient of $\log(n_{ij}/sc)$ is equal to one, which means that it must be specified as an offset.

c) The base group for gender is men and for age 0-17, and from the R-output one can see that the population is counted in 100000 individuals. Hence $\exp(\beta_0)$ is the rate of killed per 100000 in the base group, $(n_{11}/100000)\exp(\beta_0)$ is the expected number of death in this group, and $(n_{11}/100000)\exp(\hat{\beta}_0)$ is the fitted value for this combination of the factor levels.

The gender effect is estimated as $\hat{\beta}_1 = -1.0212$. The Wald statistic for the test $H_0 : \beta_1 = 1$ vs $H_1 : \beta_1 \neq 1$ is $\frac{(\hat{\beta}_1 - 1}{\hat{se}(\hat{\beta}_1)}$ where $se_{\hat{\beta}_1}(\beta_1)$ is the standard error of β_1 and $\hat{se}(\hat{\beta}_1) = se_{\hat{\beta}_1}(\hat{\beta}_1)$. From the output $se = 0.1858$, so the test statistic is $-0.0212/0.1858 = -0.1141$ and the p-value is $2P(Z > 0.1141) = 0.91$ where Z is a standard normally distributed variable, so there is no reason to reject the null hypothesis.

d) The estimated predictor for women of age 45-54 is $0.1506 - 1.0212 + 1.5366 = 0.6660$, so the estimated rate of deaths pr 10000 is $exp(0.6660) = 1.9465$. The population in this group is 3.38505×10000 so the fitted value is $3.38505\exp(0.6660) = 6.5888$ and residual is $2 - 6.5888 = -4.5888$ since the number of fatal accidents was 2.

e) The y be the vector of responses where the first 8 elements are the number of accidents for men in age group $i = 1, \ldots, 8$ and the 8 last ones are the number of accidents for women. The design matrix is then

$$X = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

The fitted values $\hat{\mu}$ satisfies the first order requirements $\frac{\partial l}{\partial \beta} = X'D(y - \hat{\mu}) = 0$ where l is the log likelihood function and $D = diag(\frac{\partial \theta_i}{\partial \eta_i})$. For the canonical link $\theta = \eta$ so $D = I_{16}$, the identity matrix of order 16. Then $X'y = X'\hat{\mu}$. The coefficient for gender is β_1 so $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{16} x_{i2}(y_i - \hat{\mu}_i) = 0$. But $x_{i2} = 0,\ i = 1, \ldots, 8$ and $x_{i2} = 1,\ i = 9, \ldots, 16$. Hence $\sum_{i=9}^{16} y_i = \sum_{i=9}^{16} \hat{\mu}_i$. The left hand side is the sum of accidents among women and the right hand side is the sum of fitted values for women.

Problem 3

a) Define the matrices

$$X_i = \begin{pmatrix} 1 & redage_{i1} \\ 1 & redage_{i2} \\ 1 & redage_{i3} \\ 1 & redage_{i4} \end{pmatrix},\ i = 1, \ldots, 5.$$

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}y_{i4})' = (\text{bone}_{i1}, \text{bone}_{i2}, \text{bone}_{i3}, \text{bone}_{i4})'$ be the responses. Then the model may be written on matrix form as

$$\mathbf{y}_i = X_i \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + Z_i \begin{pmatrix} b_{i,1} \\ b_{i,2} \end{pmatrix} + \varepsilon_i$$

where $Z_i = X_i$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$.

Here X_i is the design matrix for the fixed effects part. The random vectors $\mathbf{b}_i = (b_{i,1}, b_{i,2})'$ define the random part. The fitted values are in this case 5 non-paralell lines (random slope and intercept). The model is appropriate when it is the distribution of the intercepts and slopes which is of primary interest, not the intercept and slope for particular units.

The assumptions are that the random vectors $\mathbf{b}_i = (b_{i,1}, b_{i,2})$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ are independent and with multinormal distributions with expectation zero. The covariance matrix of $\mathbf{b}_i$ has the form $D = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix}$. The covariance of ε_i, Σ_i can be general, but is often of the form $\sigma^2 I_4$ where I_4 is a 4×4 identity matrix.

b) Since $\mathbf{b}_i$ and ϵ_i are indrpendent multinormally distributed also the distribution of $\mathbf{y}_i$ is multinormal.

The expectation of the response is $X_i\beta$ where X_i are the design matrix where the elements are the values of the covariates in cluster i.

Using that $\mathbf{b_i}$ and ϵ_i are independent the covariance matrix of the response is $V_i = Cov(\mathbf{y}_i) = Cov(Z_i\mathbf{b}_i) + Cov(\epsilon_i)$. Since $Cov(Z_i\mathbf{b_i}) = Z_iCov(\mathbf{b}_i)Z_i' = Z_iDZ_i'$ and $Cov(\epsilon_i) = \Sigma_i$, the marginal covariance matrix is $V_i = Z_iDZ_i' + \Sigma_i$.

Referring to the R-output $X_i = Z_i$ contains the measured values of the centered age of the five boys at the four occasions, i.e.

$$Z_i = \begin{pmatrix} 1 & -0.75 \\ 1 & -0.25 \\ 1 & 0.25 \\ 1 & 0.75 \end{pmatrix},$$

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)' = (52.690, 1.424)'$, $\hat{D} = \begin{pmatrix} 0.8172867^2 & 0.8172867 \times 0.732361 \\ 0.8172867 \times 0.7323611 \times 0.586 & 0.7323611^2 \end{pmatrix}$ and $\hat{\Sigma}_i = 0.2939400^2 I_4$ for $1 = 1, \cdots, 5$.

c) Since $\mathbf{y}_1, \ldots, \mathbf{y}_5$ are independent and only $\mathbf{y}_i$ is correlated with $\mathbf{b}_i$, $E[\mathbf{b}_i|\mathbf{y}_1, \ldots, \mathbf{y}_5] = E[\mathbf{b}_i|\mathbf{y}_i]$.

But $(\mathbf{b}_i, \mathbf{y}_i)'$ is multinormally distributed with expectation $(0, X_i\beta)'$ and covariance matrix

$$\begin{pmatrix} D & DZ_i' \\ Z_iD & Z_iDZ_i' + \Sigma_i \end{pmatrix}.$$

Hence

$$E[\mathbf{b}_i|\mathbf{y}_i] = E[\mathbf{b}_i] + Cov(\mathbf{b}_i, \mathbf{y}_i)[Var(\mathbf{y}_i)]^{-1}(\mathbf{y}_i - E[\mathbf{y}_i]) = DZ_i'(Z_iDZ_i' + \Sigma_i)^{-1}(\mathbf{y}_i - X_i\beta).$$

By plugging inn the REML estimates for D and Σ_i from part b) and the estimates for β from the R-output, i.e. $\hat{\beta}_0 = 52.690$ and $\hat{\beta}_1 = 1.424$, the random effects $\mathbf{b}_i$ can be estimated.

Solution proposal finals STK3100/4100-f15

Problem 1

a) The frequency function of a binomially distributed variable is

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \binom{n}{y}\exp(y\log(\pi/(1-\pi)) + n\log(1-\pi))$$

Thus $\theta = \log(\pi/(1-\pi))$, $a(\theta) = -n\log(1-\pi)$, $\phi = 1$ and $c(y,\phi) = \log\binom{n}{y}$.

The parameter θ is called the canonical parameter. The connection between the canonical parameter and the expectation is $E(y) = a'(\theta)$. If $\eta = x\beta'$ is the predictor, the link function defines the connection between the predictor and the expectation. Hence the canonical parameter can be expressed by the coefficients in the predictor, β.

b) The likelihood in a generalized linear model is $L(\theta) = \prod_{i=1}^n c(y_i,\phi)\exp(\frac{\theta_i y_i - a(\theta_i)}{\phi})$. Hence if $\check{\theta}$ and $\hat{\theta}$ are the fitted parameters in a saturated and another model the deviance Δ is -2 log likelihood ratio:

$$\Delta = 2\sum_{i=1}^n[(\check{\theta}_i - \hat{\theta}_i)y_i - a(\check{\theta}_i) + a(\hat{\theta}_i)]$$

For the binomial distribution $\check{\theta}_i = \log(y_i/(n_i - y_i))$, $\hat{\theta}_i = \log(\hat{\mu}_i/(n_i - \hat{\mu}_i))$, $a(\check{\theta}_i) = -n_i\log(1-y_i/n_i)$ and $a(\hat{\theta}_i) = -n_i\log(1-\hat{\mu}_i/n_i)$, so

$$\Delta = 2\sum_{i=1}^n[y_i\log(y_i/\hat{\mu}_i) + (n_i - y_i)\log((n_i - y_i)/(n_i - \hat{\mu}_i))]$$

The most common use of the deviance is for comparing two nested models. Then the χ^2-distribution can be a good approximation. For use of the deviance as a goodness-of-fit measure the situation is more complicated and the χ^2 approximation can be bad.

Problem 2

a) Within the same hospital $e^{\hat{\beta}_1} = 1.67$ represents the predicted proportional increase of the odds of survival of having a benign tumor (level 2) with respect to having a malign tumor.

The predicted odds for survival within country j with benign tumor is

$$\frac{\hat{\pi}_{bj}}{1-\hat{\pi}_{bj}} = \begin{cases} e^{\hat{\beta}_0+\hat{\beta}_1} & \text{if } j=1 \\ e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2} & \text{if } j=2 \\ e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_3} & \text{if } j=3 \end{cases}$$

The predicted odds for survival within country j with malign tumor is

$$\frac{\hat{\pi}_{mj}}{1-\hat{\pi}_{mj}} = \begin{cases} e^{\hat{\beta}_0} & \text{if } j=1 \\ e^{\hat{\beta}_0+\hat{\beta}_2} & \text{if } j=2 \\ e^{\hat{\beta}_0+\hat{\beta}_3} & \text{if } j=3 \end{cases}$$

Thus, the odds ratios $\text{OR} = \frac{\hat{\pi}_{bj}}{1-\hat{\pi}_{bj}} / \frac{\hat{\pi}_{mj}}{1-\hat{\pi}_{mj}} = e^{\hat{\beta}_1}$ for all three countries $j = 1, 2, 3$ or $\hat{\beta}_1 = \log \text{OR}$.

b) The output below is a deviance table from fitting various binomial models. Fill out the positions indicated by a question mark.

```
Analysis of Deviance Table

Model 1: cbind(surv, nsurv) ~ fapp + fage + fcountry
Model 2: cbind(surv, nsurv) ~ fapp + fage + finfl + fcountry
Model 3: cbind(surv, nsurv) ~ fapp + finfl + fage * fcountry
Model 4: cbind(surv, nsurv) ~ fapp * finfl + fage * fcountry
Model 5: cbind(surv, nsurv) ~ fapp * finfl + fapp * fage + fage * fcountry
Model 6: cbind(surv, nsurv) ~ fapp * finfl * fage * fcountry
  Resid. Df Resid. Dev Df Deviance
1        30     33.198
2        29     33.197  1   0.0009
3        25     25.718  4   7.4790
4        24     25.511  1   0.2079
5        22     22.059  2   3.4519
6         0      0.000 22  22.0587
```

b) Use the formula that if factor A has a levels and factor B has b levels A*B means intercept +(a-1) main effects parameters of A + (b-1) main effects parameters of B and (a-1)(b-1) interactions. Hence, remembering that the intercept and the main effects of a factor can only be counted once in a model specification:

(i) model 2 has p= 1+ 1+ 2 +1+2=7 parameters so n-p=36-7=29
(ii) model 3 has p=1+1+1+2+2+4= 11 parameters. Hence $p_{mod3} - p_{mod2} = 11 - 7 = 4$
(iii) $25.718 - 25.511 = 0.207 \approx 0.0.2079$
(iv) model 6 has 36 parameters and model 5 has 1+ 1+1+1+ +2+2 +2+4=14 parameters so $p_{mod6} - p_{mod5} = 36 - 14 = 22$.

In the remaining parts of this problem consider the hypothesis

$$H_0 : \beta_2 + \beta_3 = -1 \text{ versus } H_a : \beta_2 + \beta_3 \neq -1$$

c) $\hat{\beta}_2 + \hat{\beta}_3 + 1 = -0.6616 - 0.4946 + 1 = -0.1562$
$Var(\hat{\beta}_2 + \hat{\beta}_3 + 1) = Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2Cov(\hat{\beta}_2, \hat{\beta}_2) = 0.040 + 0.043 + 2 \times 0.021 = 0.125$ so $st.err_{\hat{\beta}_2+\hat{\beta}_3+1} = \sqrt{0.125} = 0.354$ and the Wald statistic is $-0.156/0.354 = -0.441$ which has a p-value $2P(Z \leq -0.441)) = 0.66$ for $Z \sim N(0,1)$, so the hypothesis is not rejected.

d) fcountry2 corresponds to a dummy variable, dum2, which is equal to 1 when the level of country is 2, i.e. hospital is in US, and 0 for all combinations, fcountry3 corresponds to a dummy variable, dum3, which is equal to 1 when the level of country is 3, i.e. hospital is in UK, and 0 for all combinations. Thus the model from part a) corresponds to a model $\beta_0+\beta_1 fapp+\beta_2 dum2+\beta 3 dum3$. Using that $\beta_2+\beta_3 = 1$ the model under H_0 becomes $\beta_0+\beta_1 fapp+\beta_2 dum2+(-1-\beta_2)dum3 = \beta_0+\beta_1 fapp + \beta_2(dum2 - dum3) - dum3$. This can be fitted by specifying a model of the form $offset(-dum3) + \beta_1 fapp + \beta_2(dum2 - dum3)$. Here dum2-dum3 is a variable which is 0 for treatments which takes place in Japan, 1 for treatments in US and -1 for treatments in UX. The test now consists of comparing the two deviances, and using a χ^2_1 distribution as reference.

Problem 3

a)

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i,\ i = 1, \ldots, 54$$

where

$$X_i = \begin{pmatrix} 1 & 1 & I_{[AVED\in\{7,8,9\}]} & I_{[AVED\in\{10,11,\ldots\}]} \\ 1 & 2 & I_{[AVED\in\{7,8,9\}]} & I_{[AVED\in\{10,11,\ldots\}]} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6 & I_{[AVED\in\{7,8,9\}]} & I_{[AVED\in\{10,11,\ldots\}]} \end{pmatrix}$$

$$Z_i = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 6 \end{pmatrix}$$

of dimensions 6×4 and 6×2 respectively. The indicator function is denoted as $I_{[.]}$ The fixed effects parameters are collected in the 4×1 vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_1)'$. The random effect are the elements of the 2×1 vectors $b_i = (b1_i, b2_i)', i = 1, \ldots, 54$ which is binormally distributed with expectation $(0,0)'$ and covariance matrix D and are independent of the errors $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i6})'$ where all the elements are independent $N(0, \sigma^2)$ distributed.

b) $(\hat{\beta}_1\beta_1)/\widehat{std.err}_{\hat{\beta}_1}$ is approximately $N(0,1)$ distributed which implies that an approximately 95% confidence interval has boundaries $706.00 \pm 1.9639.55$.

c) A model not containing the random effect `YEAR` is a simplifivcation of the covariance structure. This can be performed by fitting models containing `YEAR` and not containing `YEAR` by REML and comparing the values of -2 log LR. But the approximating distribution is a linear combination of χ^2-distributions, in this case $\frac{1}{1}\chi^2_1 + \frac{1}{1}\chi^2_2$.

d) The covariance matrix of y_i is $Cov(Z_i b_i + \varepsilon_i = Z_i Cov(b_i) Z_i' + \sigma^2 I_6 = Z_i D Z_i' + \sigma^2 I_6$ which equals

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 6 \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 6 \end{pmatrix}$$

$$= \begin{pmatrix} d_{11} + 2d_{12} + d_{22} & \cdots & d_{11} + 7d_{12} + 6d_{22} \\ \vdots & & \vdots \\ d_{11} + 7d_{12} + 6d_{22} & \cdots & d_{11} + 42d_{12} + 36d_{22} \end{pmatrix}$$

e) The hypothesis implies a simplification of the fixed effect structure. This can be performed by fitting the model from part a) by maximum likelihood, and also the simplified model

$$y_{ij} = \beta_0 + \beta_1 \times j + \beta_3 (AVETD_2 + 2AVETD_2) + b1_i + j \times b2_i + \varepsilon_{ij}, j = 1, \dots, 6\ , i = 1, \dots, 54$$

also by maximum likelihood. Then one compares the values of -2 log LR. The approximating distribution a χ^2_1-distribution, since the hypothesis represents one restriction.

Also a Wald test along the lines described in part 1 c) can be used. The estimate of the covariance matrix of the estimators is listed in the output.

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Monday December 1th 2014
Examination hours:	14.30 – 18.30
This problem set consists of 5 pages.	
Appendices:	None
Permitted aids:	Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Solution proposal

Problem 1

1a

i) Show that if Y is a stochastic variable with a distribution belonging to the exponential family, then $\mathrm{E}(Y) = a'(\theta)$ and $\mathrm{Var}(Y) = \phi a''(\theta)$, where a' and a'' denote the first and second derivatives of a. [Hint: Start with calculating the first derivative of $f(y;\theta,\phi)$ with respect to θ.]

Proof for $\mathrm{E}[Y] = a'(\theta)$

First derivative of f: $f'(y;\theta,\phi) = \frac{y-a'(\theta)}{\phi} f(y;\theta,\phi)$

Integral of left side:

$$\int f'(y;\theta,\phi)dy = \frac{\partial}{\partial\theta}\int f(y;\theta,\phi)dy = \frac{\partial}{\partial\theta}(1) = 0$$

Integral of right side:

$$\frac{1}{\phi}(\int y f(y;\theta,\phi)dy - a'(\theta)\int f(y;\theta,\phi)dy) = \frac{\mathrm{E}[Y] - a'(\theta)}{\phi},$$

which gives $\mathrm{E}[Y] = a'(\theta)$.

We have here assumed that differentiation and integration can be interchange.

Proof for $\mathrm{Var}(Y) = \phi a''(\theta)$:

Second derivative: $f''(y;\theta,\phi) = \left[\left(\frac{y-a'(\theta)}{\phi}\right)^2 - \frac{a''(\theta)}{\phi}\right] f(y;\theta,\phi)$

(Continued on page 2.)

Integral of left side:

$$\int f''(y;\theta,\phi)dy = \frac{\partial^2}{\partial\theta^2}\int f(y;\theta,\phi)dy = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

Integral of right side:

$$\int \left[\left(\frac{y-a'(\theta)}{\phi}\right)^2 - \frac{a''(\theta)}{\phi}\right] f(y;\theta,\phi)dy = \frac{\mathrm{Var}(Y)}{\phi^2} - \frac{a''(\theta)}{\phi}$$

which gives $\mathrm{Var}(Y) = \phi a''(\theta)$.

1b

i) Show that the Poisson distribution belongs to the exponential family.

The probability mass function can be written

$$f(y;\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda) = \frac{1}{y!}\exp(y\log(\lambda) - \lambda),$$

and it therefore belongs to the exponential family with

- $\theta = \log(\lambda)$
- $a(\theta) = \lambda = \exp(\theta)$
- $\phi = 1$
- $c(y;\phi=1) = \frac{1}{y!}$

ii) Show that $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

$\mathrm{E}[Y] = a'(\theta) = \exp(\theta) = \lambda$

$\mathrm{Var}(Y) = \phi a''(\theta) = \exp(\theta) = \lambda$

1c

i) Give an interpretation of the parameter β_1 or some transformation of it.

If we first calculate $\mu = \mu(x_1, x_2) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and then increase the value first explanatory by one and calculate $\mu' = \mu(x_1+1, x_2) = \exp(\beta_0 + \beta_1(x_1+1) + \beta_2 x_2)$, then the ratio $\mu'/\mu = \exp(\beta_1)$.

So, the expectation of the response increases by a multiplicative factor $\exp(\beta_1)$ when x_1 is increased by one unit and x_2 is unchanged.

The expectation of a Poisson distributed variable is also the rate of a Poisson process over a given time interval, so $\exp(\beta_1)$ can also be called a rate-ratio and β_1 can be called a log-rate-ratio.

ii) Assume then that $\beta_0 = 1$, $\beta_1 = 2$ and β_2=3 and predict the response Y for $x_1 = 1$ and $x_2 = 1$ and then for $x_1 = 2$ and $x_2 = 1$.

$\widehat{y} = \widehat{\mu} = \exp(1 + 2\cdot 1 + 3\cdot 1) = \exp(6) = 403.4288$

$\widehat{y} = \widehat{\mu} = \exp(1 + 2\cdot 2 + 3\cdot 1) = \exp(8) = 2980.958$

(Continued on page 3.)

1d

i) Explain what over-dispersion means in Poisson regression.

If the Poisson assumption holds, then for an observation Y_i, $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i)$. Over-dispersion occurs if $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i)$.

ii) Explain why the results above show that the current count data are over-dispersed.

The `phihat` which is reported here is an estimate of the dispersion factor ϕ, given by $\widehat{\phi} = (1/(n-p+1)) \sum_i (y_i - \widehat{\mu}_i)^2/\widehat{\mu}_i$, where n is the number of observations and $p+1$ is the number or parameters estimated. Since `phihat` is much larger than 1, the data are over-dispersed.

iii) Discuss shortly two different possibilities for performing a more correct analysis than that given above.

Possibility 1 - Quasi-likelihood: Specify only a structure on the expectation and another on the variance, but do not assume a specific distribution. For instance, assume $\mathrm{Var}(Y_i) = \phi\mu_i$. The quasi-likelihood estimates are then exactly the same as the Poisson estimates, but their standard errors are adjusted.

Possibility 1 - Negative binomial distribution: Assume that the response follows a negative binomial distribution, which is a distribution for count data that allows over-dispersion. The variance structure is then $\mathrm{Var}(Y_i) = \mu_i + \theta\mu_i^2$, where θ is a positive parameter that controls the degree of over-dispersion.

Problem 2

2a

i) Give an interpretation of a regression coefficient β, or a transformation of it, in binary regression with logit link function.

Odds is defined as $p/(1-p)$, so $g(p) = \log(p/(1-p))$ is the log-odds. Furthermore the ratio of two odds is called an odds-ratio. If we compute a probability p for some values of the explanatory variables and a probability p' for the same values of the explanatory variables except for the j-th variable, which is increased by one, then $g(p) - g(p') = \log(p/(1-p)) - \log(p'/(1-p')) = \log[p/(1-p)]/[p'/(1-p')] = \beta_j$. Therefore β_j is the log-odds-ratio and $\exp(\beta)$ is the odds-ratio for one unit increase in the j-th explanatory variable.

ii) Give then a simpler interpretation of β which holds approximately for small probabilities.

When both the probabilities p and p' are small, the odds ratio $[p/(1-p)]/[p'/(1-p')]$ is approximately p/p', which we can call a relative risk. Then $\exp(\beta_j)$ is the relative risk for one unit increase in the j-th explanatory variable.

(Continued on page 4.)

2b

i) Define AIC.

AIC = -2 log likelihood + 2 p where p is the number of parameters in the model. It gives a balance between the fit to the data and the number of parameters in the model.

ii) Which one of the models above would you choose based on the given results? Why?

The model `m1.probit` with only x_1 and the probit link has the lowest AIC value, and could therefore be chosen as the preferred model. But the model `m1.logit` with only x_1 and the logit link has almost the same AIC value. I personally think that models with logit link are easier to interpret, therefore I choose the `m1.logit` model.

2c

i) Define the two terms sensitivity and specificity.

Sensitivity: Proportion of correct predictions when true $Y_i = 1$

Specificity: Proportion of correct predictions when true $Y_i = 0$

ii) Describe what a ROC (Receiver Operating Characteristics) curve is, and draw a plot with one curve for a model with good classification performance and another model which is no better than random classification.

Compute the sensitivity and the specificity for different threshold values γ between 0 and 1. Plot sensitivity on the y-axis and (1-specificity) on the x-axis. A curve for a good model lies in the upper left corner. A curve for random classification is a straight line.

Problem 3

3a

i) Discuss whether the random effect term b_i is an important part of the model compared to other parts of the model.

The estimated variance of b_i is 0.0061 and its standard deviation is 0.0818. On the original scale, the factor $\exp(b_i)$ is between $\exp(-1.96 \cdot 0.0818) = 0.85$ and $\exp(1.96 \cdot 0.0818) = 1.18$ for 95 % of the individuals, and this can be a seen as an important difference between individuals. However, x_1 does also account for individual variation, and $\text{Var}(\beta_1 x_1)$ is estimated to be $2.06202^2 \cdot 0.087 = 0.369$ which is much higher than the variance of b_i. Other terms in the model have also much higher variance than b_i. So, compared to the rest of the model, the random term b_i can be neglected.

3b

i) Use the information you have to suggest simplifications or improvements of the model.

(Continued on page 5.)

The p value for the Wald test for $H_0 : \beta_3 = 0$ is 0.404, so x_3 can be deleted from the model. This is confirmed by the scatter plot with $\log(y)$ vs. x_3 which shows no obvious relation between $\log(y)$ and x_3.

The scatter plot of $\log(y)$ vs. x_4 shows a clear non-linear relationship between the two. One possibility can be to divide x_4 into two variables around the value of about 0.5 and estimate a model with two linear pieces joined at 0.5. Another possibility could be to constrain the curve to be flat for $x_4 > 0.5$. A third probability could be to include a second order term of x_4, i.e. x_4^2.

END

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Proposed solution: Generalized linear models
Day of examination:	Friday December 6'th 2013
Examination hours:	14.30 – 18.30

This problem set consists of 4 pages.

Appendices:	Table over normal distribution and table over χ^2-distribution
Permitted aids:	Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Problem 1

a) The expectation $\mu = E(Y) = a'(\theta)$ defines a parameterization in terms of μ. The predictor $\eta = \sum \beta_i x_i$ is connected to μ through a link function g by $\eta = g(\mu)$. The link function g must be monotone and differentiable. The dispersion parameter ϕ can be known or unknown and must then be estimated.

b) A distribution in the exponential family has a density/frequency distribution which are sufficiently regular so differentiation under integral or termwise in a sum is permitted. Differentiate once w.r.t θ in the integral $\int \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 1$ to get

$$\int \frac{(y - a'(\theta))}{\phi} \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 0$$

which simplifies to $\mu = E(Y) = a'(\theta)$. Another differentiation yields

$$\int [\frac{-a''(\theta)}{\phi} + \frac{(y - a'(\theta)^2}{\phi^2}] \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 0$$

which by using $\mu = E(Y) = a'(\theta)$ simplifies to $\phi a''(\theta) = E[(Y-\mu)^2] = Var(Y)$.

c) A saturated model is a model with a parameterization which yields the best possible fit. Then $y_i = \check{\mu}_i$, which defines $\check{\theta}_i$ and $\check{\eta}_i$ through $\mu = a'(\theta)$ and $\eta = g(\mu)$. The deviance of a model is twice the difference between the maximal log likelihood value of the saturated model and the model under consideration, i.e between

$$\sum_{i=1}^{n} \frac{\check{\theta}_i y_i - a(\check{\theta}_i)}{\phi} + c(y;\phi) \text{ and } \sum_{i=1}^{n} \frac{\hat{\theta}_i y_i - a(\hat{\theta}_i)}{\phi} + c(y;\phi)$$

(Continued on page 2.)

so the deviance is

$$\Delta = 2\sum_{i=1}^{n} \frac{(\check{\theta}_i - \hat{\theta}_i)y_i - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi}.$$

Let Mod1 and Mod 2 be two models where Mod1 is nested in Mod2. If $\hat{\hat{\theta}}_i$ and $\hat{\theta}_i$ are the estimates from Mod1 and Mod2, the difference between the deviance of Mod1 and Mod2 is $2\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \hat{\hat{\theta}}_i)y_i - a(\hat{\theta}_i) + a(\hat{\hat{\theta}}_i)}{\phi}$. The maximal value of the log likelihood under Mod1 is $\log(L_{1,max}) = \sum_{i=1}^{n} \frac{(\hat{\hat{\theta}}_i - a(\hat{\hat{\theta}}_i)}{\phi} + c(y;\phi)$ and similarly for Mod2. Hence, since the likelihood ratio is $L_{1,max}/L_{2,max}$, $-2\log(L_{1,max}/L_{2,max}) = 2\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \hat{\hat{\theta}}_i)y_i - a(\hat{\theta}_i) + a(\hat{\hat{\theta}}_i)}{\phi}$, which is the difference of the deviances.

In large samples the difference of the deviances is approximately χ^2 distributes with degrees of freedom equal to the difference of number of parameters in Mod2 and Mod1.

The deviance residuals are the square roots of the n terms, multiplied by the sign of the difference between the observed and fitted values, in the sum defining the deviance Δ.

d) The frequency function of a Poisson distributed variable is $\frac{1}{y!}\mu^y \exp(-\mu) = \exp(\log(\mu)y - \mu + \log(y!)$. Therefore, for a Poisson distributed response $\theta = \log(\mu)$ so $a(\theta) = \exp(\theta) = \mu$ and $\check{\theta}_i = \log(\check{\mu}_i) = \log(y_i)$. The dispersion parameter equals 1, so the deviance is, when $\hat{\mu}_i,\ i = 1, \ldots, n$ are the fitted values,

$$\Delta = 2\sum_{i=1}^{n}[y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)].$$

The deviance residuals are $\text{sign}(y_i - \hat{\mu}_i)\sqrt{2}[y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)]^{1/2},\ i = 1, \ldots, n$

e) The fitted values for the model with a single group is $\hat{\hat{\mu}}_i = 1.399,\ i = 1, \ldots, 679$ and 1.457 and 1.113. Hence the difference of the deviances is $2\log(1.457/1.399)(0\times128+1\times161+\cdots+6\times2)+2\log(1.113/1.399)(0\times44 + 1\times26 + \cdots + 5\times1) = 8.23$ The difference of the number of parameters is 2-1=1. The 0.99 and 0.999 quantiles in a χ^2-distribution with 1 degree of freedom are 6.64 and 10.83, so the p-value is between 0.01 and 0.001, which means a clear rejection on the 5% level since the p-value is less than 0.05.

f) For both models $E(Y_i) = \mu_i = \exp(\alpha + \beta x_i)$ where the covariate x_i is either equal to zero for both of education groups , or equal to 0 in one, and 1 in the other. The log likelihood is therefore proportional to $\sum_{i=1}^{679}[(\alpha + \beta x_i)y_i - \exp((\alpha + \beta x_i)]$. The maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ must satisfy the first order conditions. The one resulting from differentiating with respect to α is $\sum_{i=1}^{679}[y_i - \exp(\hat{\alpha} + \hat{\beta}x_i)] = 0$. Because $\exp(\hat{\alpha} + \hat{\beta}x_i) = \hat{\mu}_i$, $\sum_{i=1}^{679}(y_i - \hat{\mu}_i) = 0$.

We see that the result holds in general for models with Poisson distributed response with canonical link if the linear predictor contains

(Continued on page 3.)

a consistent term, corresponding to α in the present case. More generally it holds for all GLM with the log link if the predictor contains a constant.

Problem 2

a) On vector form for responses $\mathbf{Y}_i^T = (Y_{i1}, Y_{i2}, Y_{i3})^T$

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix} = \begin{pmatrix} 1 & \texttt{age}_i & \texttt{cyear}_{i1} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i2} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i3} & \texttt{educ}_i & \texttt{sex}_i \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b}_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{pmatrix}$$

which has the form of a linear mixed model $\mathbf{Y}_i = X_i\beta + Z_i\mathbf{b}_i + \varepsilon_i$, where the columns of the $n_i \times q$ matrix Z_i is a subset of the columns of the $n_i \times (p+1)$ design matrix X_i. Here $\mathbf{b}_i$ and ε_i are independent, $\mathbf{b}_i \sim N_q(0, D)$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{in_i})' \sim N_{n_i}(0, \Sigma_i)$, $i = 1, \ldots, N$. In this case $N = 42, n_i = 3, p = 4, q = 1$ and $\Sigma_i = \sigma^2 I_3$ where I_3 is the 3×3 identity matrix. Here $\beta_0, \ldots, \beta_4$ describe the fixed effects and $\mathbf{b}_i$ the random effects.

b) From the output we see that the coefficient of `cyear`, β_3 is estimated as $\hat{\beta} = 0.084163$ with estimated standard error 0.0081889 . Hence, an approximate 95% confidence interval has limits $0.084163 \pm 1.96 \times 0.0081889$, so the interval is (0.06811232, 0.10021281). The interval does not contain 0, so a test for constant nominal income would be clearly rejected.

c) The question involves the fixed effects. One way to do it is to fit two models one full model containing all the fixed effects and one nested model where the two effects `age` and `educ` are omitted, and obtain the maximal value of the likelihood at the two models, $L_{\max,full}$ and $L_{\max,nested}$. The likelihood ratio test consists of comparing $-2\log(L_{\max,nested}/L_{\max,full})$ to a χ^2-distribution where the degrees of freedom are the difference of the number of parameters in the two models, i.e. the number of restrictions which is two in this case. It is important that the ordinary maximum likelihood estimates are used. The restricted maximum likelihood method REML, consists of basing the estimates on a linear transformation of the data. These transformation are different for the full and nested models and involve unequal reductions of the data. Therefore it does not make sense to compare the REML likelihoods since they are based on different data.

An alternative is to use a Wald test. The approximate distribution of the estimators of the coefficients have covariance matrix $\Sigma_{\hat{\beta}} = (\sum_{i=1}^N (X_i'\Sigma_i(\hat{\theta})^{-1}X_i)^{-1}$ where $\Sigma_i(\theta)$ is the covariance matrix of $\mathbf{Y}_i$, and θ are the parameters that describe this covariance. The Wald statistic for the null hypothesis $H_0 : C\beta = R$ where C is a $r \times s$ matrix and R a $r \times 1$ known vector is $(C\hat{\beta} - R)^T(C\hat{\Sigma}_{\hat{\beta}}C^T)^{-1}(C\hat{\beta} - R)$ which is approximately χ^2 with s degrees of freedom under the null hypothesis. In this case s=2, $C = I_2$ and R=0.

(Continued on page 4.)

d) The estimates of $\mathbf{b}_i$ are based on the conditional expectations $E[\mathbf{b}_i|\mathbf{Y}_1,\ldots,\mathbf{Y}_N] = E[\mathbf{b}_i|\mathbf{Y}_i]$ since $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ are independent and $\mathbf{b}_i$ only depends on $\mathbf{Y}_i$. The simultaneous distribution of $\mathbf{b}_i$ and $\mathbf{Y}_i$ is a $q+n_i$-dimensional multinormal with expectation and covariance matrix

$$\begin{pmatrix} 0 \\ X_i\beta \end{pmatrix} \text{ and } \begin{pmatrix} D & DZ_i' \\ Z_iD & Z_iDZ_i'+\Sigma_i \end{pmatrix}.$$

It then follows from the properties of the multinormal distribution that

$$E[\mathbf{b}_i|\mathbf{Y}_i] = 0 + DZ_i'(Z_iDZ_i'+\Sigma_i)^{-1}(\mathbf{Y}_i - X_i\beta).$$

Hence $\mathbf{b}_i$ is estimated by

$$\hat{D}Z_i'(Z_i\hat{D}Z_i'+\hat{\Sigma}_i)^{-1}(\mathbf{Y}_i - X_i\hat{\beta})$$

where the estimates are the REML estimates.

e) From part d) it follows that $\hat{\Sigma}_{\mathbf{Y}_i} = Z_i\hat{D}Z_i' + \hat{\Sigma}_i$. In this case $Z_i = (1,1,1)^T$ and $\hat{\Sigma}_i = \hat{\sigma}^2 I_3$. From the output $\hat{d} = 0.1346215r$ and $\hat{\sigma}^2 = 0.747435^2$. Hence, if $\mathbf{1}_3 = (1,1,1)^T$

$$\hat{\Sigma}_{\mathbf{Y}_i} = 0.0419192^2\mathbf{1}_3\mathbf{1}_3^T + 0.747435^2 I_3 =$$

$$\begin{pmatrix} 0.0419192^2+0.7505293^2 & 0.0419192^2 & 0.0419192^2 \\ 0.0419192^2 & 0.0419192^2+0.7505293^2 & 0.0419192^2 \\ 0.0419192^2 & 0.0419192^2 & 0.0419192^2+0.7505293^2 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.5650514 & 0.001757219 & 0.001757219 \\ 0.001757219 & 0.5650514 & 0.001757219 \\ 0.001757219 & 0.001757219 & 0.5650514 \end{pmatrix}.$$

SLUTT

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære mo
Eksamensdag:	Torsdag 6. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 0 sider.	
Vedlegg:	Tabell over normal, χ^2 og t fordeling
Tillatte hjelpemidler: STK1100/STK1110 og STK2120	Godkjent kalkulator og formelsamling for

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

(a) Vi har at

$$\begin{aligned}
M_Y(t) &= \int \exp(yt)c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi})dy \\
&= \int c(y,\phi)\exp(\frac{y(\theta+t\phi) - a(\theta)}{\phi})dy \\
&= \int c(y,\phi)\exp(\frac{y(\theta+t\phi) - a(\theta+t\phi) + a(\theta+t\phi) - a(\theta)}{\phi})dy \\
&= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})(y,\phi)\exp(\frac{y(\theta+t\phi)-a(\theta+t\phi)+}{\phi})dy \\
&= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})
\end{aligned}$$

der vi har brukt at $f(y;\theta,\phi)$ integrerer seg til 1 for alle verdier av θ. Dermed blir

$$\begin{aligned}
M'_Y(t) &= \exp(\frac{a(\theta+t\phi)-a(\theta)}{\phi})a'(\theta+t\phi) \\
&= M_Y(t)a'(\theta+t\phi) \\
E[Y] &= M'_Y(0) = a'(\theta) \\
M''_Y(t) &= M_Y(t)a'(\theta+t\phi)^2 + M_Y(t)a''(\theta+t\phi)\phi \\
E[Y^2] &= M''_Y(0) = a''(\theta)^2 + \phi a''(\theta) \\
\mathrm{Var}[Y] &= \phi a''(\theta)
\end{aligned}$$

(Fortsettes på side 2.)

Alternativt kan en bruke at

$$\frac{\partial}{\partial\theta}\int_y f(y;\theta,\phi)dy = 0 \int_y \frac{1}{\phi}(y-a'(\theta))f(y;\theta,\phi)dy = 0$$

som gir $E[Y] = a'(\theta)$ og tilsvarende for varians. Dette krever at vi kan bytte om integrasjon og derivasjon, men vi har ikke diskutert de formelle kriterier for når dette er gyldig i kurset.

(b) Vi har at

$$\begin{aligned} f(y) =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\left\{-\frac{y^2-2\mu y+\mu^2}{2y\mu^2\sigma^2}\right\} \\ =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\left\{-\frac{y\frac{1}{2\mu^2}-\frac{1}{\mu}+\frac{1}{2y}}{\sigma^2}\right\} \\ =& \frac{1}{\sqrt{2\mu y^3\sigma}} \exp\{-\frac{1}{2y\sigma^2}\} \exp\left\{\frac{-y\frac{1}{2\mu^2}+\frac{1}{\mu}}{\sigma^2}\right\} \end{aligned}$$

som viser at

$$\begin{aligned} \theta =& -\frac{1}{2\mu^2} \\ a(\theta) =& \frac{1}{\mu} = -\sqrt{-2\theta} \\ \phi =& \sigma^2 \\ c(y;\phi) =& \frac{1}{\sqrt{2\mu y^3\phi}} \exp\{-\frac{1}{2y\phi}\} \end{aligned}$$

(c) Vi har at

$$\begin{aligned} E[Y] =& a'(\theta) = \frac{1}{\sqrt{-2\theta}} = \mu \\ \mathrm{Var}[Y] =& \phi a'(\theta) = \phi(-2\theta)^{-3/2} = \phi\mu^3 \end{aligned}$$

I tilfeller hvor variansstrukturen er tilnærmet kubisk som funksjon av forventning vil dette være nyttig. I tillegg er det en nyttig fordeling for responser som er positive.

Vi må ha at $\theta \leq 0$ som svarer til at $\mu \geq 0$. I tillegg må selvfølgelig $\phi \geq 0$.

(d) I dette tilfellet betyr det at hver observasjon har en forventning avhengig av forklaringsvariable gjennom

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

der g er en såkalt *link*-funksjon.

(Fortsettes på side 3.)

Devians er formelt definert som $2*(\tilde{l}-l)$ der l er log-likelihood innsatt ML estimater og $\tilde{l}$ er log-likelihood for den *mettede* modell.

Devians kan brukes for sammenlikning av modeller, der forskjell i devians svarer til likelihood ratio (på log skala).

For kjent spredningsparameter kan devians brukes til en "goodness of fit" test for å se om en gitt modell er god nok (det siste bør brukes med varsomhet).

(e) Definer $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Vi har at $\beta \rightarrow \eta_i \leftrightarrow \mu_i \leftrightarrow \theta_i$. Sammenhengen mellom μ_i og θ_i er definert gjennom fordeling. Sammenheng mellom η_i og μ_i er definert gjennom link-funksjon. Hvis link-funksjonen velges slik at $\eta_i = \theta_i$, så forenkles mye av matematikken, og log-likelihood funksjonen blir penere (konkav). Det medfører også at observert informasjon blir lik forventet informasjon. I dette tilfellet svarer det til at

$$\begin{aligned} g^{-1}(\eta) &= \mu \\ \mu &= a'(\theta) \end{aligned}$$

som gir at vi må ha

$$g^{-1}(\theta) = a'(\theta) = \frac{1}{\sqrt{-2\theta}}$$

eller $g(\mu) = -1/(2\mu^2)$

Oppgave 2

(a) Modellen kalles *random intercept and slope model*

Slike modeller er nyttige for å bygge inn korrelasjoner mellom variable som kommer fra samme individ/gruppe og der korrelasjoner avhenger av noen kovariater. De er også nyttige når vi ønsker å gjøre prediksjon for grupper der vi ikke har observasjoner.

(b) Vi har at $\boldsymbol{Y}_i$ er (multivariat) normal fordelt. Videre er

$$\begin{aligned} E[\boldsymbol{Y}_i] &= \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_i \\ \text{Var}[\boldsymbol{Y}_i] &= \boldsymbol{X}_i\boldsymbol{D}\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I} \end{aligned}$$

der $\boldsymbol{X}_i = (\mathbf{1}, \boldsymbol{x}_i)$.

Dette gir oss eksplisitte uttrykk for likelihooden noe som gjør det rimelig enkelt å bruke en numerisk optimerer for å finne ML estimater.

(c) Et problem med ML estimering er at de gir forventningsskjeve estimater for varianser. Denne skjevhetet kommer av at variansestimater

(Fortsettes på side 4.)

benytter seg av ulike residualer som må beregnes basert på estimater av regresjonskoeffisienter. REML ideen er å (lineær) transformere data slik at de transformerte data har en fordeling som ikke avhenger av regresjonskoeffisientene. Så brukes ML estimering på de transformerte data. Det er uendelig mange slike transformasjoner. Vi ønsker imidlertid å utnytte data så mye som mulig. Derfor transformeres de kun ned til dimensjon $n-p$ hvis det er p β'er. Sålenge transformasjonen har rang $n-p$, er resultatet invariant mhp hvilken transformasjon vi velger.

Fordelen med REML er altså forventningskorrigering av variansestimater. REML brukes derfor når en vil sammenlikne modeller med ulike variansstrukturer/tilfeldige effekter. REML har imidlertid lavere effisiens i forhold til ML, og ML brukes heller når en konsentrerer seg om faste effekter.

(d) Boksplottet viser klart at det er store variasjoner fra jordstykke til jordstykke som indikerer at innkludering av $b_{0,i}$ er fornuftig. Det andre plottet gir en viss indikasjon på at variasjonen endrer seg med tid, noe som kan fanges opp med $b_{1,i}$ leddet.

(e) Et mulig kriterie for modell-valg er AIC. Her vil det være henholdsvis 6, 7 og 9 parametre i modellene, som gir AIC verdier

Modell	M0	M1	M2
Loglik	273.84	40.85	5.68

som viser at modell M2 gir den klart laveste AIC verdi og dermed er å foretrekke.

En kunne alternativt brukt LR test, men da må en passe på at en tester på parameterverdier som ligger på randen av parameterrrommet. En bør da bruke en blanding av χ^2 fordelinger for å beregne P-verdier.

(f) Da alle faste effekter er såpass signifikante, er det ingen grunn til å fjerne noen av disse. Merk imidlertid at det er en svært høy korrelasjon mellom $b_{0,i}$ og $b_{1,i}$. Nå viste vi i forrige deloppgave at det var bedre å ha med begge enn bare $b_{0,i}$. En kunne imidlertid undersøke om det er hensiktsmessig å bare ha med $b_{1,i}$.

(En tilpasning med bare $b_{1,i}$ ga dog mye dårligere AIC verdi)

SLUTT

(Fortsettes på side 5.)

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære mo
Eksamensdag:	Mandag 5. desember 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 3 sider.	
Vedlegg:	Tabell over normal, χ^2 og t fordeling
Tillatte hjelpemidler: STK1100/STK1110 og STK2120	Godkjent kalkulator og formelsamling for

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

(a) Modellen kalles *random intercept model*

Slike modeller er nyttige for å bygge inn korrelasjoner mellom variable som kommer fra samme individ/gruppe.

(b) Vi har at

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + b_i\mathbf{1} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(0, \sigma^2\boldsymbol{I})$$

som gir at

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma_b^2\mathbf{1}\mathbf{1}^T + \sigma^2\boldsymbol{I})$$

Ved at vi har et eksplisitt uttrykk for den marginale fordeling for $\mathbf{Y}_i$, har vi også et eksplisitt uttrykk for likelihooden

$$L(\boldsymbol{\beta}, \sigma_b^2, \sigma^2) = \prod_{i=1}^{N} f(y_i; \boldsymbol{\beta}, \sigma_b^2, \sigma^2)$$

og denne kan så puttes inn i en numerisk optimerer for å finne ML-estimatene (evt REML estimater).

(c) Vi har

Parameter	β_1	β_1	β_2	σ_b	σ
Estimat	-4.028899	2.873710	-0.002898	0.040365	0.135060

Vi har at korrelasjonen mellom to variable fra samme fangst er

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \overset{estimert}{\approx} \frac{0.0016294}{0.0016294 + 0.0182412} = 0.082$$

(Fortsettes på side 2.)

(d) Vi har at $\hat{\boldsymbol{\beta}}$ er tilnærmet normalfordelt og dermed blir

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \log(66) + \hat{\beta}_2 \log(0.46)$$

også tilnærmet normalfordelt med varians

$$\begin{aligned}\text{Var}[\hat{\mu}] =& \text{Var}[\hat{\beta}_0] + \log(66)^2\text{Var}[\hat{\beta}_1] + \log(0.46)^2\text{Var}[\hat{\beta}_2] + \\ & 2\log(66)\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] + 2\log(0.46)\text{Cov}[\hat{\beta}_0, \hat{\beta}_2] + \\ & 2\log(66)\log(0.46)\text{Cov}[\hat{\beta}_1, \hat{\beta}_2] \\ =& 0.2009^2 = 0.04037467\end{aligned}$$

Dermed er et 95% konfidensintervall for μ lik

$$\hat{\mu} \pm 1.96\text{SE}(\hat{\mu}) = [7.619373, 8.407036]$$

Siden $L < \mu < U \Leftrightarrow \exp(L) < \exp(\mu) < \exp(U)$ blir dermed et 95% konfidensintervall for θ lik $[2037.283, 4478.465]$

(e) Strategi

(a) Start med modell med alle forklaringsvariable og så mange interaksjoner som mulig

(b) Finn optimal struktur på tilfeldige effekter.
Her bør REML brukes!

(c) Finn optimal struktur for faste effekter.
Her bør ML brukes!

(d) Presenter endelig modell med REML estimering.

Oppgave 2

(a) Vi har at den eksponensielle klasse er gitt ved

$$f(y; \theta, \phi) = c(y; \phi) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

For den binomiske fordeling har vi

$$\begin{aligned}f(y; \pi) =& \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ =& \binom{n}{y} \exp(y\log(\pi) + (n-y)\log(1-\pi)) \\ =& \binom{n}{y} \exp(y\log(\pi/(1-\pi)) + n\log(1-\pi)) \\ =& c(y; \phi)\exp(y\theta - a(\theta))\end{aligned}$$

(Fortsettes på side 3.)

med $\theta = \log(\pi/(1-\pi), \phi = 1$ og $a(\theta) = -n\log(1-\pi) = -n\log(1+\exp(\theta)), c(y;\pi) = \binom{n}{y}$.

Kanonisk link: $g(\pi) = \log(\pi/(1-\pi))$

Vi modellerer π gjennom $g(\pi) = \eta = \mathbf{x}^T\boldsymbol{\beta}$. Med kanonisk link mener vi at $\theta = \eta$, som forenkler matematikken involvert (score-likningene og informasjonsmatrisene blir enklere)

(b) AIC $= -2\log L(\hat{\theta}) + 2q$ der L er likelihood verdi innsatt estimat på de ukjente parametre θ og q er antall parametre i modellen. Dvs vi har et straffeledd for kompleksitet av modellen.

En velger den modell som har minst AIC verdi. Her blir det modellen med probit link-funksjon.

Her vil vi sammenlikne modeller som ikke er nøstet i hverandre. Dermed vil LR og Wald-type tester ikke være egnede. AIC kan imidlertid brukes mer generelt.

Her vil vi i praksis velge den modell som gir høyest likelihoodverdi da kompleksiteten (antall parametre) er de samme for alle valg av link-funksjoner.

(c) Likelihoodfunksjonen er i dette tilfellet vanskelig å beregne pga den latente tilfeldige effekten, som medfører at likelihooden er gitt som et integral. Ved å skrive integranden som $e^{g_i(\mathbf{b}_i)}$ og deretter gjøre en (2. ordens) Taylor tilnærming av g_i, får en Laplace approksimasjonen. Denne vil være betraktelig enklere å beregne og kan da optimeres for å finne estimater.

(d) Vi har at $-2LR = -2(-240.8 + 240.6) = 0.4$. Vi har her én parameter mer i modellen med tilfeldige effekter. Vi må imidlertid ta hensyn til at vi her tester en H_0 som ligger på randen av parameter-rommet. Dermed må vi bruke en mikstur av χ^2_0 og χ^2_1 for beregning av P-verdi. Dette svarer til å beregne P-verdien under χ^2_1 fordelingen og så dele på 2. Nå er $P(\chi^2_1 > 0.4) = 0.527$ som dermed gir en P-verdi på 0.2635. Dette tilsier at vi ikke har grunnlag for å forkaste H_0 i dette tilfellet.

(e) Da `log(haulsize)` har en liten z verdi (ikke signifikant) både for modellen(e) i oppgave 1 og i oppgave 2 så antyder dette at denne variabelen ikke er så viktig. Merk dog at vi kun bruker et delsett av det totale datasettet, og det kan være vi finner denne variabelen til å være signifikant hvis vi brukte det fulle datasett.

SLUTT

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK3100 — Innføring i genera..., løsningsforslag.
Eksamensdag:	Mandag 6. desember 2010
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normalfordeling og χ^2-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK1120/STK2120

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Løsningsforslag

Oppgave 1

a) Betegn respons med y og kovariatene med $x_1, \ldots, x_p$
Generell form for GLM:
Tetthet/punktsannsynlighet for respons: $f(y;\theta,\phi) = c(y,\phi)\exp(\frac{y\theta - a(\theta)}{\phi})$
Lineær prediktor: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
Link: $\eta = g(\mu); \mu = E(y)$

I dette tilfellet:

Tetthet for respons: $\pi^y(1-\pi)^{1-y} = \exp(y\log(\frac{\pi}{1-\pi}) + \log(1-\pi))$ dvs. $\theta = \log(\frac{\pi}{1-\pi}), \pi = \frac{\exp(\theta)}{1+exp(\theta)}, a(\theta) = -\log(1-\pi) = \log(1+exp(\theta)), c(y,\phi) = \phi = 1.$
Lineær prediktor: $\eta = \beta_0 + \beta_1 x$
Logit link: $\eta = g(\mu) = \log(\frac{\pi}{1-\pi}); \mu = E(y) = a'(\theta) = \frac{\exp(\theta)}{1+exp(\theta)} = \pi.$

b) $\pi(30) = \frac{\exp(\beta_0+\beta_1 30)}{1+\exp(\beta_0+\beta_1 30)}$.
Herav log odds $\log(\frac{\pi(30)}{1-\pi(30)}) = \beta_0 + \beta_1 30$, og log oddsforhold $\log(\frac{\pi(30)(1-\pi(40))}{(1-\pi(30))\pi(40)}) = \beta_1(30-40) = -10\beta_1$,
som innsatt gir oddsforholdet $OR_1 = \exp(-10 \times 0.07038) = 0.4946881$.

Et 95% konfidensintervall for log oddsforholdet er gitt ved $-10 \times \hat{\beta}_1 \pm 1.96 \times 10 \times s_{\hat{\beta}_1}$ der $s_{\hat{\beta}_1}$ er den estimerte standardfeilen til $\hat{\beta}_1$. Fra utskriften $-10 \times 0.07038 \pm 1.96 \times 10 \times 0.02667$, som gir intervallgrensene -1.226519 og -0.1811365. For oddsforholdet er konfidensintervallet derfor
(exp(-1.226519),exp(-0.1811365))=(0.293311841,0.8343214).

(Fortsettes på side 2.)

c) Den predikerte sannsynligheten er $\hat{\pi}(40) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 40)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 40)}$ som innsatt gir $\frac{\exp(-2.21358+40\times 0.07038)}{1+\exp(-2.21358+40\times 0.07038)} = 0.6460518$. Et 95% konfidensinterval for $\beta_0 + \beta_1 40$ er gitt ved $\hat{\beta}_0 + \hat{\beta}_1 40 \pm 1.96\sqrt{varest}$ der er $varest$ er estimatet for $Var(\hat{\beta}_0 + 40\hat{\beta}) = Var(\hat{\beta}_0) + 40^2 Var(\hat{\beta}) + 2 \times 40 \times Cov(\hat{\beta}_0, \hat{\beta}_1)$, dvs $0.99874^2 + 1600 \times 0.02667^2 + 2 \times 40 \times (-0.906) \times 0.99874 \times 0.02667 = 0.2039967$. Konfidensintervallet for $\beta_0 + \beta_1 40$ er derfor $-2.21358 + 40 \times 0.07038 \pm 1.96 \times \sqrt{0.2039967}$ eller (-0.2835082, 1.486966). Siden $\exp(x)/(1 + \exp(x)$ er en voksende funksjon blir konfidensintervallet for den predikerte sannsynligheten $\exp(-0.2835082)/(1 + \exp(-0.2835082), \exp(1.486966)/(1 + \exp(1.486966) = (0.4295939, 0.8156225)$.

d)
```
Analysis of Deviance Table

Model 1: sore ~ 1
Model 2: sore ~ I(duration)
Model 3: sore ~ I(duration) + factor(type)
Model 4: sore ~ I(duration) + I(duration^2) + factor(type)
  Resid. Df Resid. Dev Df Deviance
1        34     46.180
2        33     33.651  1   12.528
3        32     30.138  1    3.513
4        31     30.133  1    0.005
```

Her ser vi at en test for model 3 mot model 4, dvs H_0 at leddet x^2 kan sløyfes ikke er signifikant, p-verdien er 0.95 i en χ^2 fordeling med 1 frihetsgrad. Heller ikke neste forenkling, dvs at hypotesen at koeffesienten for faktoren `type` er lik null, er spesielt signifikant, p-verdien er 0.06=P(X>3.513) i en χ^2 fordeling med 1 frihetsgrad. Ytterligere forenklinger gir derimot sterk signifikans, p-verdien er 0.0004=P(X>12.528) i en χ^2 fordeling med 1 frihetsgrad.

e) Ved biomisk respons er deviansen

$$2\sum_{i=1}^{n}[y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})]$$

der y_i er observerte og $\hat{\mu}_i$ er tilpassede verdier.

Ved binær respons er $n_i = 1, i = 1, \cdots, n = 35$ og $\hat{\mu}_i = \hat{\pi}_i$ slik at deviansen blir

$$2\sum_{i=1}^{35}[y_i \log(\frac{y_i}{\hat{\pi}_i}) + (1 - y_i)\log(\frac{1 - y_i}{1 - \hat{\pi}_i})].$$

Siden y_i har verdiene 0 eller 1, blir uttrykk av formen $y_i \log(y_i)$ og $(1 - y_i)\log(1 - y_i)$ lik 0. Deviansen reduseres derfor til

$$-2\sum_{i=1}^{35}[y_i \log(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}) + \log(1 - \hat{\pi}_i)].$$

(Fortsettes på side 3.)

Ligningene for bestemmelse av SME er $X'D(y-\hat{\mu})=0$, jfr ligning (5.4) på side 68 i læreboka, der X er designmatrisen, y er vektoren av observasjoner og $\hat{\mu}=\hat{\pi}$ er vektoren av tilpassede verdier. Matrisen D er diagonalmatrisen med diagonalelementer av formen $1/V(\hat{\mu}_i)g'(\hat{\mu}_i), i=1,\cdots,n$, der V er variansfunksjonen og g linkfunksjonen. I dette tilfellet reduseres D derfor til identitetsmatrisen slik at ligningene blir $X'y=X'\hat{\pi}$.

Da er den første summen i deviansen $-2y'\hat{\eta}$ der $\eta=X\hat{\beta}$ er vektoren av lineære prediktorer. Derfor er $-2y'\hat{\eta}=-2y'X\hat{\beta}=-2(X'y)'\hat{\beta}=-2(X'\hat{\pi})'\hat{\beta}=-2\hat{\pi}'X\hat{\beta}=-2\hat{\pi}'\hat{\eta}$, som gir resultatet.

Vi ser at deviansen bare avhenger av observasjonene gjennom de tilpassede verdiene. Det går derfor ikke an å sammenligne de observerte vardiene og de tilpassede verdiene dvs. vurdere føyningen ved å se på deviansen.

Oppgave 2

a) Antallet dødsfall av denne typen kan også ses som antallet suksesser"(i dette tilfellet dødsfall), i et stort antall forsøk. Hvert personår er et forsøk. Det betyr at antallet suksesser er binomisk fordelt. Her er suksess-sannsynheten, p, liten og antallet forsøk, m, stort, og da er fordelingen til antallet tilnærmet Poissonfordelt med forventing λ når mp er nær λ.

Forventet antall kan uttrykkes som en rate som er proposjonal med størrelsen på området, intervallet eller populasjonen antallet angis for, dvs $\lambda=N\mu$ der N er størrelsen og μ raten. I dette tilfellet angis størrelsen med antall personår.

Modellen blir derfor at responsene $y_1,\cdots,y_n$ er uavhengige Poisson-fordelte variable med forventning $N_i\mu_i$ der N_i er antallet personår i gruppene og μ_i er ratene.

I modellen som er tilpasset angis alder med et andregradspolynom, og røyking er en faktor med to nivåer. I tillegg er det et samspillsledd for koeffesienten foran 1'te gradsleddet og faktoren røyking. Det betyr at raten har formen

$$\mu=\begin{cases}\exp(\beta_0+\beta_1 age+\beta_2 age^2) & \text{for ikke}-\text{roeykere}\\ \exp(\beta_0+\beta_3+(\beta_1+\beta_4)age+\beta_2 age^2) & \text{for roeykere}\end{cases}$$

Vi ser fra modeltilpasningen at både alder og røyking ser ut til å ha betydning, men samspillsleddet gjør sammenhengen mer komplisert.

b) Siden fornetningen utrykkes som $N_i\mu_i=\exp(\log(N_i))\mu_i$ vil den lineære prediktoren i tillegg til den lineære kombinasjonen beskrevet ovenfor, inneholde et ledd av typen $\log(N_i)$, med andre ord et ledd hvor koeffesienten er kjent og lik 1. Ledd av denne typen, der koeffesientene ikke skal estimeres, betegnes som "offset".

c) Fra uttykket i punkt a) ser vi at forholdet mellom ratene for røykere og ikke-røykere blir

$$\frac{\exp(\beta_0+\beta_3+(\beta_1+\beta_4)age+\beta_2 age^2)}{\exp(\beta_0+\beta_1 age+\beta_2 age^2)}=\exp(\beta_3+\beta_4 age)$$

(Fortsettes på side 4.)

som altså varierer med alderen. For leger på 40 år er det estimerte forholdet $\exp(\hat{\beta}_3 + \hat{\beta}_4 \times 40) = \exp(2.370 - 0.03084 \times 40) = 3.114493$ og leger på 70 $\exp(\hat{\beta}_3 + \hat{\beta}_4 \times 70) = \exp(2.370 - 0.03084 \times 70) = 1.234766$. Dødeligheten på grunn av hjertesykdommer som kan tilskrives røyking, er altså nesten en tredjedel for 70 åringer i forhold til 40 åringer. En forklaring kan rett og slett være at de som er mest utsatt på grunn av røking har en overdødelighet i yngre alder.

d) Betrakt nullhypotesen $H_0 : C\beta = r$ der C er en $q \times 6$ matrise, r er en $q \times 1$ vektor og $\beta' = (\beta_0, \beta_1, \cdots, \beta_5)$ er vektoren av ukjente parametre. Wald resten gir forkastning for store verdier av $(C\hat{\beta} - r)'(C\Sigma_{\hat{\beta}}C')^{-1}(C\hat{\beta} - r)$ der $\hat{\beta}$ er sannsynlighetsmaksimeringsestimatoren og $\Sigma_{\hat{\beta}}$ er den estimarte kovariansmatrisen til $\hat{\beta}$. Under H_0 er testobservatoren tilnærmet χ^2_q-fordelt. I dette tilfellet er $H_0 : \beta_4 = \beta_5 = 0$, som svarar til $q = 2$, $r = (0, 0)'$. C matrisen har rader $(0, 0, 0, 0, 1, 0)$ og $(0, 0, 0, 0, 0, 1)$ slik at $C\Sigma_{\hat{\beta}}C'$ er den estimerte kovariansmatrisen til $(\hat{\beta}_4, \hat{\beta}_5)$, altså den som er oppgitt i oppgaveteksten. Da er Wald-observatoren

$$(-0.0975518230, 0.0005195636) \begin{pmatrix} 1.143363e-02 & -8.807653e-05 \\ -8.807653e-05 & 6.844424e-07 \end{pmatrix}^{-1} \begin{pmatrix} -0.0975518230 \\ 0.0005195636 \end{pmatrix}$$

som er lik 9.85051 og gir klar forkastning ved sammenligning med en χ^2-fordeling med 2 frihetsgrader.

e) I modellene ovenfor brukes kanonisk link, dvs $\eta = \log(\mu) = \theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ slik at likelihooden er

$$\prod_{i=1}^{n} \exp(y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))$$

og log-likelihood

$$l = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})].$$

Herav

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} [y_i x_{ij} - x_{ij} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]$$

og

$$\frac{\delta^2 l}{\delta \beta_j \delta \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

for alle $j, k = 0, \ldots, p$, der $x_{i0} = 1, i = 1, \ldots, n$. Den observerte informasjonsmatrisen har elementer $-\frac{\delta^2 l}{\delta \beta_j \delta \beta_k}$, som alle er ikke-tilfeldige. De forventede verdiene blir derfor de samme. Siden den forventede informasjoensmatrisen er forventningen av den observerte, må de i dette tilfellet bli like.

SLUTT

STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 1 of 2

Submission deadline

Thursday October 2 2025, 14:30 in Canvas (canvas.uio.no).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts. You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with Latex). The assignment must be submitted as **a single PDF file**. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons,you have to contact the person responsible for the course, Ingrid Hobæk Haff (e-mail: ingrihaf@math.uio.no) no later than the same day as the deadline. All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Specifically about this assignment

In order to get the assignment accepted you need to fulfil the following requirements:

- Made a real attempt on all (sub-)questions
- Give satisfactory answers in at least 60% of the (sub-)questions
- Include relevant code outputs in your report.

Complete guidelines about delivery of mandatory assignments:

www.uio.no/english/studies/examinations/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1

In this problem, we will consider the relationship between a person's wingspan and height. The wingspan is the horizontal measurement from fingertip to fingertip with outstretched arms. The data below show the wingspan and height in cm for 16 Australian women, and are also recorded in the file http://www.uio.no/studier/emner/matnat/math/STK3100/data/wingspan.txt:

	Height (x)	Wingspan (y)
1	63.0	62.0
2	63.0	62.0
3	65.0	64.0
4	64.0	64.5
5	68.0	67.0
6	69.0	69.0
7	71.0	70.0
8	68.0	72.0
9	68.0	70.0
10	72.0	72.0
11	73.0	73.0
12	73.5	75.0
13	70.0	71.0
14	70.0	70.0
15	72.0	76.0
16	74.0	76.5

We will return to the relationship between height and wingspan in question f), but first we will consider the problem more generally. To this end we consider a simple linear regression model with the single covariate $\mathbf{x} = (x_1, \ldots, x_n)^T$ and the response $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma^2)$.

We will consider two models, M_0 and M_1. Model M_1 is the standard linear regression model

$$\mu_i = \beta_0 + \beta_1 x_i,$$

whereas M_0 is the same model, but without the intercept, i.e.

$$\mu_i = \beta_1^* x_i.$$

Now, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ be the $n \times 1$ vector of mean values. Further, the model matrices for models M_0 and M_1 are denoted $\mathbf{X}_0$ and $\mathbf{X}_1$ and the model

spaces are denoted $C(\mathbf{X}_0)$ and $C(\mathbf{X}_1)$. We use the notation $\mathbf{1}_k$ to denote a $k \times 1$ vector of 1s.

a) Give the model matrices for models M_0 and M_1. What are the ranks of the two model matrices? Explain what it means that the models are nested.

b) Give the projection matrices $\mathbf{P}_0$ and $\mathbf{P}_1$ onto the two model spaces.

c) Use the projection matrices to show that the vectors of fitted values for models M_0 and M_1 may be given, respectively, as

$$\hat{\boldsymbol{\mu}}_0 = \hat{\beta}_1^* \mathbf{x}$$

and

$$\hat{\boldsymbol{\mu}}_1 = \bar{Y}\mathbf{1}_n + \hat{\beta}_1(\mathbf{x} - \bar{x}\mathbf{1}_n),$$

with

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

d) Show that

$$\mathbf{Y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}/\sigma^2 \quad \text{and} \quad \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{Y}/\sigma^2$$

are independent and determine their distributions (Hint: Use Cochran's theorem). It is sufficient to determine the distributions under M_0.

e) Show that the F -statistic for testing the null hypothesis that model M_0 holds versus the alternative hypothesis that model M_1 holds may be given as

$$F = \frac{\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0\|^2}{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}_1\|^2/(n-2)} = \frac{\sum_{i=1}^n \left(\bar{Y} - \hat{\beta}_1 \bar{x} + \left(\hat{\beta}_1 - \hat{\beta}_1^*\right) x_i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y} - \hat{\beta}_1(x_i - \bar{x})\right)^2 /(n-2)},$$

and determine the distribution of the F-statistic under model M_0.

We now return to the example on wingspan and height, considered in the beginning of the problem.

f) Read the data in the table given in the beginning of the problem into R. Use the `lm` command to fit models M_0 and M_1, and use the `anova` command to perform the F test. Discuss your results.

Problem 2

In this problem, we will consider how the survival of a passenger on the Titanic depended on ticket class and the passenger's age. The file

`https://www.uio.no/studier/emner/matnat/math/STK3100/data/titanic.txt`

consists of three columns with the following information about a subset of 70 passengers:

- `survived`: survived the shipwreck (0 = no, 1 = yes)
- `age`: age of the passenger in years
- `pclass`: ticket class, grouped as either 3rd class or 1st/2nd class (0 = 1st/2nd class, 1 = 3rd class)

We will use logistic regression to analyse the data using R. You may read the data into R by the commands:

```
data="http://www.uio.no/studier/emner/matnat/math/STK3100/data/titanic.txt"
titanic=read.table(data,header=T)
```

a) Fit a logistic regression model with `survived` as response and `pclass` as the only covariate. Is there a significant effect of `pclass`?

b) Denote by β_1 the regression coefficient for `pclass` in the logistic regression model. Give an interpretation of e^{β_1}, and estimate it.

c) Fit a logistic regression model where you also include `age` as a covariate. Denote by β_2 the regression coefficient for `age` in this model. Give an interpretation of e^{β_2}, and estimate it.

d) Use the Wald test, the likelihood ratio test and the score test to test the hypothesis that $\beta_2 = 0$ in the model in question c). How well do the tests agree? What are the conclusions of the tests?

e) Find 95% confidence intervals for e^{β_1} from b) and for e^{β_2} from c). Give interpretations of these intervals.

STK3100/4100—Introduction to Generalized Linear Models

Mandatory assignment 2 of 2

Submission deadline

Thursday November 6 2025, 14:30 in Canvas (canvas.uio.no).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts. You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with Latex). The assignment must be submitted as **a single PDF file**. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons,you have to contact the person responsible for the course, Ingrid Hobæk Haff (e-mail: ingrihaf@math.uio.no) no later than the same day as the deadline. All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Specifically about this assignment

In order to get the assignment accepted you need to fulfil the following requirements:

- Made a real attempt on all (sub-)questions
- Give satisfactory answers in at least 60% of the (sub-)questions
- Include relevant R outputs in your report.

Complete guidelines about delivery of mandatory assignments:

www.uio.no/english/studies/examinations/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1

In this problem, we will look at data from a fertility study in Botswana, where the aim is to investigate how the number of living children a woman gives birth to depends on whether she has education, uses contraception, lives in an urban area, etc.

You may read the data into R by the commands:

```
data_file <- "http://www.uio.no/studier/emner/matnat/math/STK3100/data/fertility_data.csv"
fertility <- read.csv(data_file,header=TRUE)
```

The data file consists of 8 columns with the following variables:

- `educ0`: indicator of whether the woman has education (0 = no; 1 = yes)
- `usemeth`: indicator of whether the woman uses contraception (0 = no; 1 = yes)
- `urban`: indicator of whether the woman lives in an urban area (0 = no; 1 = yes)
- `electric`: indicator of whether the woman has electricity installed in her house (0 = no; 1 = yes)
- `radio`: indicator of whether the woman has a radio (0 = no; 1 = yes)
- `tv`: indicator of whether the woman has a TV (0 = no; 1 = yes)
- `bicycle`: indicator of whether the woman has a bicycle (0 = no; 1 = yes)
- `ceb`: number of living children.

In this problem, we will assume that the number of living children (`ceb`) is Poisson distributed within each of the $2^7 = 128$ combinations of the 7 binary covariates.

a) Expain why this may be a reasonable assumption.

b) Fit a GLM for Poisson data with logarithmic link function to the data, using all the covariates.

c) Perform an analysis that clarifies the significance of the different covariates, where you consider the effect of removing some of the covariates from the model. Which of the models you have considered seems to give the best description of the data?

d) Interpret the estimates from "the best model" in question c) as rate ratios, and give 95% confidence intervals for the rate ratios.

e) Estimate the claim rate of a woman who does not have education, does not use contraception, does not live in an urban area, has electricity, a radio and a bicycle, but no TV. Also give a 95% confidence interval for this rate.

Problem 2

The Poisson distribution has variance equal to the mean. In practice this assumption is often unrealistic for count data, because the variability is in fact greater than can be described by the Poisson mean. This is what we call *overdispersion*. A common way to handle overdispersed count data is to use a type of mixture of Poisson distributions, which results in the negative binomial distribution. In this problem we will consider some properties of the negative binomial distribution and the corresponding GLMs. As shown in the lectures, the negative binomial distribution may be obtained as a mixture of Poisson distributions.

More specifically, if Λ is a random variable that follows the gamma distribution with pdf

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

and further, the random variable Y, given $\Lambda=\lambda$, is Possion distributed with parameter λ, and thus has the conditional pmf

$$p(y|\lambda)=\frac{\lambda^y}{y!}e^{-\lambda},\quad y=0,1,2,\ldots$$

Then, the marginal pmf of Y is given by

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

which is the pmf of the negative binomial distribution.

We will now assume that $k>0$ is a given constant, and consider the random variable $Y^*=Y/k$. Then $P(Y^*=y^*)=P(Y=ky^*)$, for $y^*=0,\frac{1}{k},\frac{2}{k},\ldots$, so Y^* has pmf

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

a) Show that (2) is a distribution in the exponential dispersion family. That is, show that (2) can be written on the form $\exp((\theta y^*-b(\theta))/a(\phi)+c(y;\phi))$, with $a(\phi)=1/k$, and determine θ and $b(\theta)$.

b) Find the mean and variance of Y^* using the relations (4.3) and (4.4) in the text book. Use these results to show that $\mathrm{E}(Y)=\mu$ and determine $\mathrm{Var}(Y)$.

Then we assume that $Y_1, \ldots, Y_n$ are independent and have pmf of the form (1), and that their means $\mu_i = \mathrm{E}(Y_i)$ are specified via a link function g, i.e. $g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}$.

c) Derive an expression for the log-likelihood function $L(\boldsymbol{\mu}, k; \mathbf{y})$. (In the text book, there is an expression of the log-likelihood for the parameterisation with $\gamma = 1/k$. You should express it in terms of k.)

d) For a given $k > 0$, the deviance for a negative binomial GLM is given by $D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2(L(\mathbf{y}, k; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}, k; \mathbf{y}))$. Derive an expression for $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$.

e) Derive the limit of the deviance when $k \to \infty$. How can you explain this result?

We will now return to the fertility data from Problem 1, where it was assumed that the Poisson distribution was a good fit, and hence, that there was no overdispersion.

f) Fit your prefered GLM from Problem 1 c), substituting the Poisson distribution with the negative binomial (this is done using the command *glm.nb* from the `MASS` package, see the R code on the horseshoe crab data from the lecture on October 21). Does it provide a better fit than the Poisson GLM? What does the estimated k (called θ in the R output) tell you about possible over-dispersion, and how do you see that in light of your response to Problem 1 a)?

Formulas in STK3100/4100

1) Linear models and least squares

a) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, then the least squares estimate $\widehat{\boldsymbol{\beta}}$ and the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})$.

b) Let $C(\boldsymbol{X})$ denote the model space, i.e. the subspace of $\mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{X}$, and let $\boldsymbol{P}_{\boldsymbol{X}}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$), and $\mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{trace}(\boldsymbol{P}_{\boldsymbol{X}})$.

c) The projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, i.e. it depends only on the subspace $C(\boldsymbol{X})$ and not on the choice of basis vectors for the subspace. If $\boldsymbol{X}$ has full rank, we have $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

d) For a random vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ and a fixed matrix $\boldsymbol{A}$, we have $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, written $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, if its joint pdf is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp\{-(1/2)(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

b) Suppose $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ is partitioned as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

then

$$\boldsymbol{Y}_1|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

c) [Cochran's theorem] Assume that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k$ are projection matrices with $\sum_{i=1}^k \boldsymbol{P}_i = \boldsymbol{I}$. Then $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}$ are independent for $i = 1, \ldots k$, and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{Y}/\sigma^2$ has a non-central chi-squared distribution with non-centrality parameter $\lambda_i = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\mu}/\sigma^2$ and degrees of freedom equal to the rank of $\boldsymbol{P}_i$.

3) Generalized linear models (GLMs)

a) A random variable Y_i has a distribution in the exponential dispersion family if its pmf/pdf may be written

$$f(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\},$$

where θ_i is the natural parameter and ϕ is the dispersion parameter. We have $E(Y_i) = b'(\theta_i)$ and $\text{var}(Y_i) = b''(\theta_i)a(\phi)$.

b) For a GLM we have that $Y_1, \ldots Y_n$ are independent with pmf/pdf from the exponential dispersion family. The linear predictors $\eta_1, \ldots, \eta_n$ are given by $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i\boldsymbol{\beta}$, and the expected values $\mu_i = E(Y_i)$ satisfy $g(\mu_i) = \eta_i$ for a strictly increasing and differentiable link function g. For the canonical link function $g(\mu_i) = (b')^{-1}(\mu_i)$ we have $\theta_i = \eta_i$.

c) The likelihood equations for a GLM are given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for} \quad j = 1, \ldots, p.$$

d) Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood (ML) estimator for a GLM. Then

$$\widehat{\boldsymbol{\beta}} \overset{approx}{\sim} N\left(\boldsymbol{\beta}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\right)$$

where $\boldsymbol{X}$ is the model matrix and $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i)$.

e) Consider a GLM with $a(\phi) = \phi/\omega_i$. Let $\widehat{\mu}_i = b'(\widehat{\theta}_i)$ be the ML estimate of μ_i under the actual model, and let $y_i = b'(\tilde{\theta}_i)$ be the ML estimate of μ_i under the saturated model. Then

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

where

$$D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \omega_i \left[y_i\left(\tilde{\theta}_i - \widehat{\theta}_i\right) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i)\right]$$

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ for $i = 1, \ldots n$ are independent vectors that correspond to d observations from each of n clusters. A normal linear mixed effects model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots \epsilon_{id})^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u}_i$. Often one will have $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$.

b) For a generalized linear mixed model we assume that the conditional pmf/pdf of Y_{ij} given $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ is in the exponential dispersion family, and that for a link function g we have

$$g\left[E(Y_{ij} \mid \boldsymbol{u}_i)\right] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i.$$