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)

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

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

b

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

c

Assume that $\mathbf{y} = (y_1, \dots, y_n)^T$ is an observed value of the random vector $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, and let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$. 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

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

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/preplanned)
- `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)
```

```
    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 $\boldsymbol{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{V} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}}\}$. We consider the linear model $\boldsymbol{\mu} = \boldsymbol{X\beta}$, where the model matrix $\boldsymbol{X}$ is a $n \times p$ matrix, and assume that $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$. If we observe $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \dots, y_n)^{\mathrm{T}}$, 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 $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{AY}) = \mathrm{trace}(\boldsymbol{AV}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ 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 $\boldsymbol{P}_1, \dots, \boldsymbol{P}_k$ 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 $i = 1, \dots k$, 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 $Y_1, \dots Y_n$ 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

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

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

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