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

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

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 $V_1, \dots, V_N$ are independent with $V_i \sim \text{bin}(n_i, \pi_i)$, $i = 1, \dots, N$, and let $Y_i = V_i/N_i$. Let $x_1, \dots, x_N$ denote known explanatory variable values. Consider a generalized linear model (GLM) for $Y_1, \dots, Y_N$, 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 $\boldsymbol{P_X}$ denote the projection matrix onto $C(\boldsymbol{X})$. Then $\widehat{\boldsymbol{\mu}} = \boldsymbol{P_X}\boldsymbol{y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P_X}^2 = \boldsymbol{P_X}$), and $\mathrm{rank}(\boldsymbol{P_X}) = \mathrm{trace}(\boldsymbol{P_X})$.

c) The projection matrix $\boldsymbol{P_X}$ 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 $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

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

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