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

$$\begin{aligned} P(Y=y) = \exp\{\log P(Y=y)\} &= \exp\left\{\log\binom{n}{ny} + ny\log\pi + (n-ny)\log(1-\pi)\right\} \\ &= \exp\left\{\log\binom{n}{ny} + ny\log\frac{\pi}{1-\pi} + n\log(1-\pi)\right\} \end{aligned}$$

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)

$$\mathrm{Var}(Y) = a(\phi)b''(\theta) = \frac{1}{n}\frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{\left(1+e^\theta\right)^2} = \frac{1}{n}\frac{e^\theta}{\left(1+e^\theta\right)^2} = \frac{1}{n}\pi(1-\pi)$$

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)

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1$$

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 $\frac{\partial \mu_i}{\partial \eta_i} = b''(\theta_i) = \frac{e_i^\theta}{(1+e_i^\theta)^2} = \pi_i(1-\pi_i)$, and $\mathrm{var}(Y_i) = \frac{1}{n_i}\pi_i(1-\pi_i)$, 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)

$$w_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\mathrm{var}(Y_i)} = \frac{(\pi_i(1-\pi_i))^2}{\frac{1}{n_i}\pi_i(1-\pi_i)} = n_i\pi_i(1-\pi_i)$$

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 $\boldsymbol{\mu} = \boldsymbol{X\beta}$, 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 $\widehat{\boldsymbol{\mu}} = \boldsymbol{P_X y}$. The projection matrix is symmetric and idempotent (i.e. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P_X}$), 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 $E(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{AY}) = \mathrm{trace}(\boldsymbol{AV}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A\mu}$.

2) Multivariate normal distribution and normal linear models

a) $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ 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

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