

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 exponential dispersion family $E(Y) = b'(\theta)$ and $\text{Var}(Y) = a(\phi)b''(\theta)$. 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

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

(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:

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

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

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

Hence, the likelihood equations are given by

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

which is equivalent to

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

d

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

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

where $\hat{\theta}_i = \frac{1}{\hat{\mu}_i^2}$, $b(\hat{\theta}_i) = \frac{2}{\hat{\mu}_i}$ 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

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

e

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

(Continued on page 3.)

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

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 $\hat{\phi}_{MLE} = \hat{\sigma}_{MLE}$ is given by

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

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

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

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

Problem 2

a

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

- $Y_1, \dots, Y_N$ are independent with $Y_i \sim \text{bin}(\pi_i, 1)$, $i = 1, \dots, N$
- the linear predictor is given by $\eta_i = \mathbf{x}_i \boldsymbol{\beta}$, $i = 1, \dots, N$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{i8})$ with $x_{i1} = 1$, $x_{i2}, \dots, x_{i6}$ 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 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_8)^T$
- the mean $\mu_i = E(Y_i)$ is related to the linear predictor via the link function as $\eta_i = g(\mu_i) = \text{logit}(\mu_i) = \mu_i / (1 - \mu_i)$.

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 has $p_3 = 8$ 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 straightforward 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 $\Phi^{-1}(P(Y_{ij} = 1|u_i)) = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i$,

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

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

$$E(Z - u_i) = E(Z) - E(u_i) = 0 - 0 = 0$$

$$\text{Var}(Z - u_i) \stackrel{\text{ind.}}{=} \text{Var}(Z) + (-1)^2 \cdot \text{Var}(u_i) = 1 + \sigma_u^2,$$

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

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

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 individuals, 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 independent. 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 $\hat{\sigma}_u^2 = 2.242$, which we can use to compute $\hat{\beta}_j / \sqrt{1 + \hat{\sigma}_u^2}$, $j = 1, \dots, 7$, and it turns out that these are approximately equal to the estimated β -coefficients of the probit model from **Problem 2c**, which is in line with the result from **a**. Further, if we look at the P-values from the Wald tests for the β -coefficients, we see that several of the effects that were significant in the GLM, more specifically **Sex** and **Weight**, are no longer significant. 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.