

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)

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

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 $E(Y) = \mu$ and to find $\text{Var}(Y)$.

Assume in the rest of the problem that the random variables $Y_1, \dots, Y_n$ are independent with $Y_i \sim IG(\mu_i, \sigma)$, $i = 1, \dots, n$. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the vectors of covariate values, with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$. 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

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

d

Show that the deviance of this GLM is given by

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

e

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

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**: Weight 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 , $i = 1, \dots, n$, $j = 1, \dots, d$, and $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})$ be vectors of explanatory variables. A binomial generalised linear mixed model (GLMM) with probit link and random intercept $u_i \sim N(0, \sigma_u^2)$, $i = 1, \dots, n$, is then given by

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

(Continued on page 5.)

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

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

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

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

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

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 $\hat{\sigma}_u^2 = 2.242$).
- 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 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ be a vector of random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ and covariance matrix $\mathbf{V} = E\{(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T\}$. We consider the linear model $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where the model matrix $\mathbf{X}$ is a $n \times p$ matrix, and assume that $\mathbf{V} = \sigma^2 \mathbf{I}$. If we observe $\mathbf{Y} = \mathbf{y} = (y_1, \dots, y_n)^T$, then the least squares estimate $\hat{\boldsymbol{\beta}}$ and the fitted values $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are obtained by minimizing $\|\mathbf{y} - \boldsymbol{\mu}\|^2 = (\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{y} - \boldsymbol{\mu})$.

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

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

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

2) Multivariate normal distribution and normal linear models

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

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

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

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

then

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

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

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 $\eta_i = \sum_{j=1}^p x_{ij}\beta_j = \mathbf{x}_i\boldsymbol{\beta}$, 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

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

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

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}), \quad \text{approximately}$$

where $\mathbf{X}$ is the model matrix and $\mathbf{W}$ 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 $\hat{\mu}_i = b'(\hat{\theta}_i)$ 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

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

where

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

is the deviance.

4) Normal and generalized linear mixed models

a) We assume that $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{id})^T$ 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

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

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 $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{id})^T \sim N(\mathbf{0}, \mathbf{R})$ is independent of $\mathbf{u}_i$. Often one will have $\mathbf{R} = \sigma^2 \mathbf{I}$.

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

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