UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100/STK4100 — Introduction to generalized linear models.

Day of examination: Wednesday December 18th 2019

Examination hours: 9.00 – 13.00.

This problem set consists of 0 pages.

Appendices: Formulas for STK3100 and STK4100

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

a) The odds is defined as

$$\text{Odds}(x_i) = \frac{\pi_i}{1-\pi_i} = \frac{\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}}{\frac{1}{1+\exp(\alpha+\beta x_i)}} = \exp(\alpha+\beta x_i)$$

Then we can write an odds-ratio between observations with explanatory variables $x_i' = x_i + 1$ and x_i as

$$\text{OR} = \frac{\text{Odds}(x_i+1)}{\text{Odds}(x_i)} = \frac{\exp(\alpha+\beta(x_i+1))}{\exp(\alpha+\beta x_i)} = \exp(\beta)$$

When both π_i and $\pi_{i'}$ are small then $1-\pi_i \approx 1$ and $1-\pi_{i'} \approx 1$. Thus $\text{Odds}(x_i) \approx \pi_i$ and $\exp(\beta) = \text{OR} \approx \pi_{i'}/\pi_i = \text{RR}$, i.e. a relative risk

b) We have

$$\text{P}(Y_i = 1|Z_i = 1) = \frac{\text{P}(Y_i = 1 \cap Z_i = 1)}{\text{P}(Z_i = 1)}$$

where $\text{P}(Y_i = 1 \cap Z_i = 1) = \text{P}(Z_i = 1|Y_i = 1)\text{P}(Y_i = 1) = \rho_1\pi_i$.

Similarily $\text{P}(Y_i = 0 \cap Z_i = 1) = \text{P}(Z_i = 1|Y_i = 0)\text{P}(Y_i = 0) = \rho_0(1-\pi_i)$.

Thus $\text{P}(Z_i = 1) = \rho_1\pi_i + \rho_0(1-\pi_i)$ and

$$\text{P}(Y_i = 1|Z_i = 1) = \frac{\rho_1\pi_i}{\rho_1\pi_i + \rho_0(1-\pi_i)} = \frac{\rho_1\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}}{\rho_1\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)} + \rho_0\frac{1}{1+\exp(\alpha+\beta x_i)}}$$

(Continued on page 2.)

which simplifies to

$$P(Y_i = 1|Z_i = 1) = \frac{\rho_1 \exp(\alpha + \beta x_i)}{\rho_1 \exp(\alpha + \beta x_i) + \rho_0} = \frac{\exp(\alpha^* + \beta x_i)}{1 + \exp(\alpha^* + \beta x_i)}$$

with $\alpha^* = \alpha + \log(\rho_1/\rho_0)$.

The implication of the result is that the same odds-ratio $\exp(\beta)$ can be estimated both on cohort (population) and on case-control data.

Problem 2

a) The gamma density can be rewritten as

$$\begin{aligned} f(y;\mu,k) &= \exp(-\tfrac{k}{\mu}y - k\log(\mu) + k\log(k) + (k-1)\log(y) - \log(\Gamma(k))) \\ &= \exp(\tfrac{(-1/\mu)y - \log(\mu)}{1/k} - k\log(k) + (k-1)\log(y) - \log(\Gamma(k))) \\ &= \exp((\theta y - b(\theta))/\phi + c(y,\phi)) \end{aligned}$$

with $\theta = -1/\mu, b(\theta) = \log(\mu) = -\log(1/\mu) = -\log(-\theta), \phi = 1/k$ and $c(y,\phi) = -\log(1/\phi)/\phi + (1/\phi - 1)\log(y) - \log(\Gamma(1/\phi))$.

We have $E[Y] = b'(\theta) = -\frac{1}{-\theta}(-1) = -\frac{1}{\theta} = \mu$ and $\text{var}[Y] = \phi b''(\theta) = \phi\frac{1}{\theta^2} = \phi\mu^2$.

b) A GLM consists of three components

(i) Independent Y_i from a distribution $\exp((\theta_i y - b(\theta_i))/\phi + c(y,\phi))$ with $\mu_i = b'(\theta_i)$

(ii) A linear predictor $\eta_i = \sum_{j=1}^p \beta_j x_{ij}$

(iii) A link function $g(\mu_i) = \eta_i$.

We saw in a) that (i) was satisfied, so a GLM for gamma distributed Y_i thus requires specification of (ii) the linear predictor η_i and (iii) the link function $g()$.

The log-likelihood can be written as $L(\beta) = \sum_{i=1}^n (y_i\theta_i - b(\theta_i))/\phi + c(y_i,\phi)$. By the chain rule the derivatives of $L(\beta)$ then becomes

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \eta_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial(y_i\theta_i - b(\theta_i))}{\partial \theta_i}\frac{1}{\phi} = \sum_{i=1}^n x_{ij}\frac{\partial \mu_i}{\partial \eta_i}\frac{(y_i - \mu_i)}{\phi}\frac{\partial \theta_i}{\partial \mu_i}$$

and since $\frac{\partial \theta_i}{\partial \mu_i} = 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/b''(\theta_i)$ and $\text{var}[Y_i] = \phi b''(\theta) = \phi\mu_i^2$ for the gamma family we obtain

$$\sum_{i=1}^n \frac{y_i - \mu_i}{\phi\mu_i^2} x_{ij}\frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for } j = 1,\dots,p$$

(Continued on page 3.)

c) Since $\mathrm{E}[Y_i] - \mu_i = 0$ as long as the μ_i are correctly specified the score equations are still unbiased. The construction is called the (score equations) for quasi-likelihood which only require the structure of expectation and variance to be correctly specified to give asymptotically normal estimators with expected information matrix equal to the covariance matrix of the scores.

The dispersion parameter ϕ can be estimated using the Pearson X^2 with the particular variance structure, thus we can use

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$$

as a consistent estimator of ϕ when $\mathrm{var}[Y_i] = \phi\mu_i^2$.

d) We see that the price depends significantly on x_1 size, x_4 rent, x_5 distance to the east (x) and also to whether the appartement has a balcony (x_3), but that the number of rooms (x_2) does not a significantly effect (when adjusting for size). Increase of x_1 and x_3 increases the price, whereas increase in x_4 and x_5 decreases the price.

The natural estimate for $\mu = \eta = \beta_0 + \sum_{j=1}^{5} \beta_j x_j = \beta^t \mathbf{x}$ equals $\hat{\mu} = \hat{\beta}_0 + \sum_{j=1}^{5} \hat{\beta}_j x_j = 526.64 + 18.4 * 70 + 25.77 * 2 - 0.12745 * 1000 - 93.29 * 2 =$ 1552 x 1000 NOK.

With $\hat{\mu} = \hat{\beta}'\mathbf{x}$ we can estimate $\mathrm{var}(\hat{\mu})$ by $\mathbf{x}^t\hat{\Sigma}\mathbf{x}$ where $\hat{\Sigma}$ is the estimated covariance matrix for $\hat{\beta}$.

e) The linear regression model used here can be expressed as

$$Y_i = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \varepsilon_i$$

where the $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ and independent.

Roughly the $\hat{\beta}_j$ and p-values correspond well between the gamma-regression and the usual linear regression model. The residual plots of (deviance) residuals vs. fitted values reveal no clear non-linearities for either model. However, it seems that the residuals for the linear regression model tend to increase as the fitted values increases. It appears that for the gamma model the deviance residuals does not display such a tendency. It may thus be that the gamma model with variance structure $\phi\mu^2$ captures heteroscedasticity better than the constant variance in the linear regression.

(Although the homoscedastic model did not affect estimates or p-values much it will certainly be an important issue if we want prediction intervals for predicted prizes with a given new vector of explanatory variables $\mathbf{x}$. But this perspective was not discussed seriously in STK3100/4100 this semester).

(Continued on page 4.)

Problem 3

a) We have $\mathrm{E}[Y_{ij}] = \mathrm{E}[\mathrm{E}(Y_{ij}|u_i)] = \mathrm{E}[\exp(\beta_0 + \beta_1 x_{ij} + u_i)] = \exp(\beta_0 + \beta_1 x_{ij})\mathrm{E}[e^{u_i}]$ since $\beta_0 + \beta_1 x_{ij}$ is a constant, not a random variable.

Also, $\mathrm{E}[e^{u_i}] = M(1) = \exp(\sigma_u^2/2)$, thus marginally $\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$ and only the intercept β_0 is changed relative to the generalized linear mixed effects model.

b) We have that $\mathrm{var}[Y_{ij}] = \mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|u_i)]$. Here $\mathrm{var}(Y_{ij}|u_i)] = \exp(\beta_0 + \beta_1 x_{ij} + u_i) = \mathrm{E}(Y_{ij}|u_i)]$ and so $\mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$.

Furthermore, $\mathrm{var}[\mathrm{E}(Y_{ij}|u_i)] = \mathrm{var}(\exp(\beta_0 + \beta_1 x_{ij} + u_i)) = \exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{var}(\exp(u_i)) = \exp(2(\beta_0 + \beta_1 x_{ij})\ [M(2) - M(1)^2]$ and $M(2) - M(1)^2 = \exp(2\sigma_u^2) - \exp(\sigma_u^2) = \exp(\sigma_u^2)(\exp(\sigma_u^2) - 1)$. Thus

$$\mathrm{var}[Y_{ij}] = \exp(\beta_0+\beta_1 x_{ij}+\sigma_u^2/2)+\exp(2(\beta_0+\beta_1 x_{ij}))\exp(\sigma_u^2)(\exp(\sigma_u^2)-1).$$

c) Also when Y_{ij} given u_i is gamma distributed we get $\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2)$ by the same derivation as in a). The result that only the intercept changes from the mixed to the marginal model depends only on the log-link structure.

For the marginal variances of the Y_{ij} we again have $\mathrm{var}[Y_{ij}] = \mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|u_i)]$ and

$$\begin{aligned}\mathrm{var}[\mathrm{E}(Y_{ij}|u_i)] &= \mathrm{var}(\exp(\beta_0 + \beta_1 x_{ij} + u_i)) \\ &= \exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{var}(\exp(u_i)) \\ &= \exp(2(\beta_0 + \beta_1 x_{ij}))\exp(\sigma_u^2)(\exp(\sigma_u^2) - 1)\end{aligned}$$

as in question b).

Furthermore, $\mathrm{var}(Y_{ij}|u_i) = \phi\mu_{ij}^2 = \phi\exp(2(\beta_0 + \beta_1 x_{ij} + u_i))$, so

$$\begin{aligned}\mathrm{E}[\mathrm{var}(Y_{ij}|u_i)] &= \mathrm{E}[\phi\exp(2(\beta_0 + \beta_1 x_{ij} + u_i))] \\ &= \phi\exp(2(\beta_0 + \beta_1 x_{ij}))\mathrm{E}[\exp(2u_i)] \\ &= \phi\exp(2(\beta_0 + \beta_1 x_{ij}))M(2) = \phi\exp(2(\beta_0 + \beta_1 x_{ij}) + 2\sigma_u^2)\end{aligned}$$

As in b) the answer is then obtained adding these two terms.

END

(Continued on page 5.)