UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

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

Day of examination: Wednesday December 2nd 2020

Examination hours: 9.00 – 13.00.

This problem set consists of 5 pages.

Appendices: Formulas in STK3100 / STK4100

Permitted aids: All resources

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

Problem 1

a) We rewrite $\frac{\lambda^y}{y!}\exp(-\lambda) = \exp(y\log(\lambda) - \lambda - \log(y!))$ which gives $\theta = \log(\lambda))$, $\lambda = \exp(\theta) = b(\theta)$ and $c(y) = -\log(y!)$.

By general results $\mu = \mathrm{E}[Y] = b'(\theta) = \exp(\theta) = \lambda$ and $\mathrm{var}[Y] = b''(\theta) = \exp(\theta) = \lambda = \mu$.

b) For $y = 1, 2, 3, \dots$ we have

$$\mathrm{P}(Y = y|Y > 0) = \frac{\mathrm{P}(Y = y)}{\mathrm{P}(Y > 0)} = \frac{\mathrm{P}(Y = y)}{1 - \mathrm{P}(Y = 0)} = \frac{\lambda^y \exp(-\lambda)/y!}{1 - \exp(-\lambda)}$$

Then rewrite $\frac{\lambda^y}{y!}\exp(-\lambda)/(1-\exp(-\lambda)) = \exp(y\log(\lambda) - \lambda - \log(y!) - \log(1 - \exp(-\lambda)))) = \exp(y\theta - b(\theta) + c(y))$ with $\theta = \log(\lambda)$ (as in a)), $b(\theta) = \lambda + \log(1 - \exp(-\lambda)) = \exp(\theta) - \log(1 - \exp(-\exp(\theta)))$ and $c(y) = -\log(y!)$ (also as in a)).

c) We have (when $f(y;\gamma)$ is a density, otherwise replace integral by sum) $\mathrm{P}(Y \in B) = \int_B f(y;\gamma)dy = \exp(-b_0(\gamma))\int_B \exp(\gamma y - c_0(y))dy$. Thus $Y|Y \in B$ has a density

$$f_B(y;\theta) = \frac{f(y;\gamma)}{\mathrm{P}(Y\in B)} = \frac{\exp(\gamma y - b_0(\gamma) + c_0(y))}{\exp(-b_0(\gamma))\int_B \exp(\gamma y - c_0(y))dy}$$

$$= \exp(\gamma y - \log(\textstyle\int_B \exp(\gamma y - c_0(y))dy) + c_0(y))$$

and so $\theta = \gamma$, $c(y) = c_0(y)$ and $b(\theta) = \log(\int_B \exp(\theta y - c(y))dy)$.

(Continued on page 2.)

Problem 2

a) The logistic regression model here is

$$\mathrm{P}(Y_i = 1|x_{i1}, x_{i2}) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2}}.$$

Thus we get the odds

$$\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2})} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2})$$

which lead to the *odds-ratio*

$$(\frac{\mathrm{P}(Y_i = 1|x_{i1}+1, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}+1, x_{i2}))})/(\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}))}) = \exp(\beta_1)$$

as general interpretations of $\exp(\beta_1)$ as *odds-ratios* when changing x_{i1} by one unit keeping x_{i2} constant and similarly for $\exp(\beta_2)$

When all $\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})$ are small we have $1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}) \approx 1$ and so

$$(\frac{\mathrm{P}(Y_i = 1|x_{i1}+1, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}+1, x_{i2}))})/(\frac{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})}{(1 - \mathrm{P}(Y_i = 1|x_{i1}, x_{i2}))}) \approx \frac{\mathrm{P}(Y_i = 1|x_{i1}+1, x_{i2})}{\mathrm{P}(Y_i = 1|x_{i1}, x_{i2})},$$

i.e. as a *relative risk*.

Inserting estimates $\hat{\beta}_j$ leads to estimated odds-ratios approximated by estimated relative risks. Here we get $\exp(\hat{\beta}_1) = \exp(2.20) = 9.02$ so as an approximation bad health increases the chance of frequent doctorial visits by a factor 9 (Actually this will be an exaggerated increase since $exp(2.20) = 9.02$ is a large value). Similarly $\exp(\hat{\beta}_2) = \exp(-0.338) = 0.71$, so after the health reform the proportions of women with frequent doctoral visits were approximately reduced by 30%.

b) Approximately the MLE $\hat{\beta}_j \sim \mathrm{N}(\beta_j, se_j^2)$ (by slight abuse of notation since se_j are statistics/random variables) and so

$$\begin{aligned} 0.95 &\approx \mathrm{P}(-1.96 < (\hat{\beta}_j - \beta_j)/se_j < 1.96) \\ &= \mathrm{P}(\hat{\beta}_j - 1.96 se_j < \beta_j < \hat{\beta}_j + 1.96 se_j) \\ &= \mathrm{P}(\exp(\hat{\beta}_j - 1.96 se_j) < \exp(\beta_j) < \exp(\hat{\beta}_j + 1.96 se_j)) \end{aligned}$$

Inserting the estimated regression coefficients and standard errors gives 95% confidence interval

$(6.51, 12.51)$ for $\exp(\beta_1)$

$(0.52, 0.98)$ for $\exp(\beta_2)$,

(Continued on page 3.)

none of which overlaps the value 1. Thus we can reject both null hypotheses H_{0j} at a 5 percent level. In particular the interval for $\exp(\beta_1)$ has a low end far from 1, indicating strong statistical significance. This is confirmed by the $|z_j| = |\hat{\beta}_j/se_j|$ being values larger than 2 and p-values less than 0.05 (in particular for $j = 1$).

c) Deviances are two times the difference between the log-likelihood with a specific model and the log-likelihood with a saturated model where fitted values $\tilde{y}_i$ are equal to observed values y_i.

Differences in deviances between two nested models, i.e. a smaller model is a special case of a larger, is chi-square distributed with degrees of freedom equal to the diiference in number of parameters between the model given that the smaller model is true.

The approximation to the χ^2 distribution stems for the differences in deviances being equal to twice the differences in log-likehoods betweens the models, as the saturated log-likelihood terms cancels out, and so is due to the properties of the likelihood ratio test.

A deviance table gives deviances, changes in deviances and changes in no. of parameters for a series of nested models. This gives the opportunity to test a series of models and evaluate which (often categorical) explanatory variables that are essential or non-essential for the outcome.

In the table below the questions marks have been replaced by the acutal numbers:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2226	1303.4	
badh	1	158.613	2225	1144.8	< 2.2e-16
reform	1	4.404	2224	1140.4	0.035849
educat	2	2.339	2222	1138.0	0.310536
inccat	2	8.641	2220	1129.4	0.013292
badh:reform	1	1.458	2219	1127.9	0.227285
badh:inccat	2	0.851	2217	1127.1	0.653313
educat:inccat	4	13.689	2213	1113.4	0.008357

Problem 3

a) We get

$$\mu = \mathrm{E}[Y] = \mathrm{E}[\exp(V)] = M_V(1) = \exp(\gamma * 1 + \frac{1}{2}\sigma^2 * 1^2) = \exp(\gamma + \frac{1}{2}\sigma^2).$$

Thus

$$\begin{aligned}\mathrm{var}[Y] &= \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 = M_V(2) - M_V(1)^2 = \exp(2\gamma + 2\sigma^2) - \exp(2\gamma + \sigma^2) \\ &= \mu^2(\exp(\sigma^2) - 1) = \phi\mu^2\end{aligned}$$

(Continued on page 4.)

with $\phi = \exp(\sigma^2) - 1$.

b) Since $\mu_i = \exp(\alpha + \beta x_i + \frac{1}{2}\sigma^2) = \exp(\gamma_i + \frac{1}{2}\sigma^2)$ with $\gamma_i = \alpha + \beta x_i$ we get that

$$\mathrm{E}[V_i] = \mathrm{E}[\log(Y_i)] = \gamma_i = \alpha + \beta x_i$$

and so (α, β) can be estimated by least squares estimates $(\hat{\alpha}, \hat{\beta})$ of a simple linear regression on $V_i = \log(Y_i)$. Furthermore σ^2 can then be estimated as $\hat{\sigma}^2 = \sum_{i=1}^n (V_i - \hat{\alpha} - \hat{\beta} x_i)^2/(n-2)$ which then leads to estimate $\hat{\phi} = \exp(\hat{\sigma}^2) - 1$ of ϕ.

c) The score equations for generalized linear models can be written as

$$\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{Y_i - \mu_i}{\mathrm{var}(Y_i)} = \frac{1}{\phi} \sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{Y_i - \mu_i}{\nu^*(\mu_i)} = 0; \quad j = 1, \ldots, p$$

for models with $\mathrm{var}(Y_i) = \phi \nu^*(\mu_i)$.

These estimating equations can also be used as so-called quasi-likelihood equations under the weaker assumption that $Y_i; i = 1, \ldots, n$, are independent, but not necessarily from an exponential dispersion family, $g(\mu_i) = g(\mathrm{E}[Y_i]) = \beta' x_i$ for a link function $g()$ and variance specification $\mathrm{var}(Y_i) = \phi \nu^*(\mu_i)$. This leads to consistent and asymptotically normal estimates of β with a variance matrix as the inverse of the information matrix (- expected Jacobi) based on the quasi-score function.

In this particular case one obtains the estimates by specifying a gamma family with an identity link in the `glm`-command (since the variance function for the gamma family equals μ^2) or equivalently by specifying a quasi-family with identity link and μ^2 variance.

Problem 4

a) When $u_i \sim \mathrm{N}(0, \sigma^2)$ it has a density $f(u; \sigma_u^2) = \exp(-u^2/(2\sigma_u^2))/\sqrt{2\pi\sigma_u^2}$. Thus the marginal probability is by the rule of double expectation

$$\mathrm{P}(Y_{ij} = 1 | x_{ij}) = \mathrm{E}[\mathrm{P}(Y_{ij} = 1 | x_{ij}, u_i)] = \int \frac{\exp(\beta_0 + \beta_1 x_{ij} + u)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u)} f(u; \sigma_u^2) du$$

Similarly, for $j = 0, 1$ and $k = 0, 1$,

$$\begin{aligned}
\pi_i(j, k; \beta_0, \beta_2, \sigma_u^2) &= \mathrm{P}(Y_{i1} = j, Y_{i2} = k | x_{i1}, x_{i2}) \\
&= \mathrm{E}[\mathrm{P}(Y_{i1} = j, Y_{i2} = k | x_{i1}, x_{i2}, u_i)] \\
&= \mathrm{E}[\mathrm{P}(Y_{i1} = j | x_{i1}, x_{i2}, u_i) \mathrm{P}(Y_{i2} = k | x_{i1}, x_{i2}, u_i)] \\
&= \int \frac{\exp(j(\beta_0 + \beta_1 x_{i1} + u))}{1 + \exp(\beta_0 + \beta_1 x_{i1} + u)} \frac{\exp(k(\beta_0 + \beta_1 x_{i2} + u))}{1 + \exp(\beta_0 + \beta_1 x_{i2} + u)} f(u; \sigma_u^2) du
\end{aligned}$$

(Continued on page 5.)

which can not be written as $\mathrm{P}(Y_{i1}=1|x_{i1})\mathrm{P}(Y_{i2}=1|x_{i2})$. Thus Y_{i1} and Y_{i2} are marginally dependent.

The marginal likelihood can then be written

$$l(\beta_0,\beta_1,\sigma_u^2)=\prod_{i=1}^n \pi_i(Y_{i1},Y_{i2};\beta_0,\beta_1,\sigma_u^2).$$

b) If $Y_{i1}+Y_{i2}=2$ then necessarily both $Y_{i1}=1$ and $Y_{i2}=1$, thus $\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=2)=1$. Similarly, $Y_{i1}+Y_{i2}=0$ imply that both $Y_{i1}=0$ and $Y_{i2}=0$ and so also $\mathrm{P}(Y_{i1}=0|Y_{i1}+Y_{i2}=0)=1$. Thus no such pair (Y_{i1},Y_{i2}) will conditionallly on $Y_{i1}+Y_{i2}$ contain any information on β_1.

But when $Y_{i1}+Y_{i2}=1$ then either $Y_{i1}=1$ and $Y_{i2}=0$ or $Y_{i1}=0$ and $Y_{i2}=1$ and so, conditionally on u_i, x_{i1} and x_{2i},

$$\begin{aligned}\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1) &= \frac{\mathrm{P}(Y_{i1}=1,Y_{i2}=0)}{\mathrm{P}(Y_{i1}=1,Y_{i2}=0)+\mathrm{P}(Y_{i1}=0,Y_{i2}=1)}\\ &= \frac{\exp(\beta_1(x_{i1}-x_{i2}))}{1+\exp(\beta_1(x_{i1}-x_{i2}))}\end{aligned}$$

since (same conditioning on u_i, x_{i1}, x_{i2})

$$\mathrm{P}(Y_{i1}=1,Y_{i2}=0)=\frac{\exp(\beta_0+\beta_1x_{i1}+u_i)}{1+\exp(\beta_0+\beta_1x_{i1}+u_i)}\frac{1}{1+\exp(\beta_0+\beta_1x_{i2}+u_i)}.$$

The expression for $\mathrm{P}(Y_{i1}=0,Y_{i2}=1)$ has the same denominator which then cancel out in $\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1)$. One is then left with

$$\begin{aligned}\mathrm{P}(Y_{i1}=1|Y_{i1}+Y_{i2}=1) &= \frac{\exp(\beta_0+\beta_1x_{i1}+u_i)}{\exp(\beta_0+\beta_1x_{i1}+u_i)+\exp(\beta_0+\beta_1x_{i2}+u_i)}\\ &= \frac{\exp(\beta_1x_{i1})}{\exp(\beta_1x_{i2})+\exp(\beta_1x_{i2})}=\frac{\exp(\beta_1(x_{i1}-x_{i2}))}{1+\exp(\beta_1(x_{i1}-x_{i2}))}\end{aligned}$$

This means that it is possible to estimate β_1 by running a logistic regression

- with outcome Y_{i1}
- for pairs with $Y_{i1}+Y_{i2}=1$
- with explanatory variables $x_{i1}-x_{i2}$
- and no intercept

END