UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100/STK4100 — Proposed solution: Generalized linear models

Day of examination: Friday December 6'th 2013

Examination hours: 14.30 – 18.30

This problem set consists of 4 pages.

Appendices: Table over normal distribution and table over χ^2-distribution

Permitted aids: Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Problem 1

a) The expectation $\mu = E(Y) = a'(\theta)$ defines a parameterization in terms of μ. The predictor $\eta = \sum \beta_i x_i$ is connected to μ through a link function g by $\eta = g(\mu)$. The link function g must be monotone and differentiable. The dispersion parameter ϕ can be known or unknown and must then be estimated.

b) A distribution in the exponential family has a density/frequency distribution which are sufficiently regular so differentiation under integral or termwise in a sum is permitted. Differentiate once w.r.t θ in the integral $\int \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 1$ to get

$$\int \frac{(y - a'(\theta))}{\phi} \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 0$$

which simplifies to $\mu = E(Y) = a'(\theta)$. Another differentiation yields

$$\int [\frac{-a''(\theta)}{\phi} + \frac{(y - a'(\theta)^2}{\phi^2}] \exp(\frac{\theta y - a(\theta)}{\phi}) c(y;\phi) dy = 0$$

which by using $\mu = E(Y) = a'(\theta)$ simplifies to $\phi a''(\theta) = E[(Y-\mu)^2] = Var(Y)$.

c) A saturated model is a model with a parameterization which yields the best possible fit. Then $y_i = \check{\mu}_i$, which defines $\check{\theta}_i$ and $\check{\eta}_i$ through $\mu = a'(\theta)$ and $\eta = g(\mu)$. The deviance of a model is twice the difference between the maximal log likelihood value of the saturated model and the model under consideration, i.e between

$$\sum_{i=1}^{n} \frac{\check{\theta}_i y_i - a(\check{\theta}_i)}{\phi} + c(y;\phi) \text{ and} \sum_{i=1}^{n} \frac{\hat{\theta}_i y_i - a(\hat{\theta}_i)}{\phi} + c(y;\phi)$$

(Continued on page 2.)

so the deviance is

$$\Delta = 2\sum_{i=1}^{n} \frac{(\check{\theta}_i - \hat{\theta}_i)y_i - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi}.$$

Let Mod1 and Mod 2 be two models where Mod1 is nested in Mod2. If $\hat{\hat{\theta}}_i$ and $\hat{\theta}_i$ are the estimates from Mod1 and Mod2, the difference between the deviance of Mod1 and Mod2 is $2\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \hat{\hat{\theta}}_i)y_i - a(\hat{\theta}_i) + a(\hat{\hat{\theta}}_i)}{\phi}$. The maximal value of the log likelihood under Mod1 is $\log(L_{1,max}) = \sum_{i=1}^{n} \frac{(\hat{\hat{\theta}}_i - a(\hat{\hat{\theta}}_i)}{\phi} + c(y;\phi)$ and similarly for Mod2. Hence, since the likelihood ratio is $L_{1,max}/L_{2,max}$, $-2\log(L_{1,max}/L_{2,max}) = 2\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \hat{\hat{\theta}}_i)y_i - a(\hat{\theta}_i) + a(\hat{\hat{\theta}}_i)}{\phi}$, which is the difference of the deviances.

In large samples the difference of the deviances is approximately χ^2 distributes with degrees of freedom equal to the difference of number of parameters in Mod2 and Mod1.

The deviance residuals are the square roots of the n terms, multiplied by the sign of the difference between the observed and fitted values, in the sum defining the deviance Δ.

d) The frequency function of a Poisson distributed variable is $\frac{1}{y!}\mu^y \exp(-\mu) = \exp(\log(\mu)y - \mu + \log(y!)$. Therefore, for a Poisson distributed response $\theta = \log(\mu)$ so $a(\theta) = \exp(\theta) = \mu$ and $\check{\theta}_i = \log(\check{\mu}_i) = \log(y_i)$. The dispersion parameter equals 1, so the deviance is, when $\hat{\mu}_i$, $i = 1, \ldots, n$ are the fitted values,

$$\Delta = 2\sum_{i=1}^{n}[y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)].$$

The deviance residuals are $\text{sign}(y_i - \hat{\mu}_i)\sqrt{2}[y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)]^{1/2}$, $i = 1, \ldots, n$

e) The fitted values for the model with a single group is $\hat{\hat{\mu}}_i = 1.399$, $i = 1, \ldots, 679$ and 1.457 and 1.113. Hence the difference of the deviances is $2\log(1.457/1.399)(0\times 128 + 1\times 161 + \cdots + 6\times 2) + 2\log(1.113/1.399)(0\times 44 + 1\times 26 + \cdots + 5\times 1) = 8.23$ The difference of the number of parameters is 2-1=1. The 0.99 and 0.999 quantiles in a χ^2-distribution with 1 degree of freedom are 6.64 and 10.83, so the p-value is between 0.01 and 0.001, which means a clear rejection on the 5% level since the p-value is less than 0.05.

f) For both models $E(Y_i) = \mu_i = \exp(\alpha + \beta x_i)$ where the covariate x_i is either equal to zero for both of education groups , or equal to 0 in one, and 1 in the other. The log likelihood is therefore proportional to $\sum_{i=1}^{679}[(\alpha + \beta x_i)y_i - \exp((\alpha + \beta x_i)]$. The maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ must satisfy the first order conditions. The one resulting from differentiating with respect to α is $\sum_{i=1}^{679}[y_i - \exp(\hat{\alpha} + \hat{\beta}x_i)] = 0$. Because $\exp(\hat{\alpha} + \hat{\beta}x_i) = \hat{\mu}_i$, $\sum_{i=1}^{679}(y_i - \hat{\mu}_i) = 0$.

We see that the result holds in general for models with Poisson distributed response with canonical link if the linear predictor contains

(Continued on page 3.)

a consistent term, corresponding to α in the present case. More generally it holds for all GLM with the log link if the predictor contains a constant.

Problem 2

a) On vector form for responses $\mathbf{Y}_i^T = (Y_{i1}, Y_{i2}, Y_{i3})^T$

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix} = \begin{pmatrix} 1 & \texttt{age}_i & \texttt{cyear}_{i1} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i2} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i3} & \texttt{educ}_i & \texttt{sex}_i \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b}_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{pmatrix}$$

which has the form of a linear mixed model $\mathbf{Y}_i = X_i\beta + Z_i\mathbf{b}_i + \varepsilon_i$, where the columns of the $n_i \times q$ matrix Z_i is a subset of the columns of the $n_i \times (p+1)$ design matrix X_i. Here $\mathbf{b}_i$ and ε_i are independent, $\mathbf{b}_i \sim N_q(0, D)$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{in_i})' \sim N_{n_i}(0, \Sigma_i)$, $i = 1, \ldots, N$. In this case $N = 42, n_i = 3, p = 4, q = 1$ and $\Sigma_i = \sigma^2 I_3$ where I_3 is the 3×3 identity matrix. Here $\beta_0, \ldots, \beta_4$ describe the fixed effects and $\mathbf{b}_i$ the random effects.

b) From the output we see that the coefficient of `cyear`, β_3 is estimated as $\hat{\beta} = 0.084163$ with estimated standard error 0.0081889 . Hence, an approximate 95% confidence interval has limits $0.084163 \pm 1.96 \times 0.0081889$, so the interval is (0.06811232, 0.10021281). The interval does not contain 0, so a test for constant nominal income would be clearly rejected.

c) The question involves the fixed effects. One way to do it is to fit two models one full model containing all the fixed effects and one nested model where the two effects `age` and `educ` are omitted, and obtain the maximal value of the likelihood at the two models, $L_{\max,full}$ and $L_{\max,nested}$. The likelihood ratio test consists of comparing $-2\log(L_{\max,nested}/L_{\max,full})$ to a χ^2-distribution where the degrees of freedom are the difference of the number of parameters in the two models, i.e. the number of restrictions which is two in this case. It is important that the ordinary maximum likelihood estimates are used. The restricted maximum likelihood method REML, consists of basing the estimates on a linear transformation of the data. These transformation are different for the full and nested models and involve unequal reductions of the data. Therefore it does not make sense to compare the REML likelihoods since they are based on different data.

An alternative is to use a Wald test. The approximate distribution of the estimators of the coefficients have covariance matrix $\Sigma_{\hat{\beta}} = (\sum_{i=1}^N (X_i'\Sigma_i(\hat{\theta})^{-1} X_i)^{-1}$ where $\Sigma_i(\theta)$ is the covariance matrix of $\mathbf{Y}_i$, and θ are the parameters that describe this covariance. The Wald statistic for the null hypothesis $H_0 : C\beta = R$ where C is a $r \times s$ matrix and R a $r \times 1$ known vector is $(C\hat{\beta} - R)^T (C\hat{\Sigma}_{\hat{\beta}} C^T)^{-1} (C\hat{\beta} - R)$ which is approximately χ^2 with s degrees of freedom under the null hypothesis. In this case s=2, $C = I_2$ and R=0.

(Continued on page 4.)

d) The estimates of $\mathbf{b}_i$ are based on the conditional expectations $E[\mathbf{b}_i|\mathbf{Y}_1,\ldots,\mathbf{Y}_N] = E[\mathbf{b}_i|\mathbf{Y}_i]$ since $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ are independent and $\mathbf{b}_i$ only depends on $\mathbf{Y}_i$. The simultaneous distribution of $\mathbf{b}_i$ and $\mathbf{Y}_i$ is a $q+n_i$-dimensional multinormal with expectation and covariance matrix

$$\begin{pmatrix} 0 \\ X_i\beta \end{pmatrix} \text{ and } \begin{pmatrix} D & DZ_i' \\ Z_iD & Z_iDZ_i' + \Sigma_i \end{pmatrix}.$$

It then follows from the properties of the multinormal distribution that

$$E[\mathbf{b}_i|\mathbf{Y}_i] = 0 + DZ_i'(Z_iDZ_i' + \Sigma_i)^{-1}(\mathbf{Y}_i - X_i\beta).$$

Hence $\mathbf{b}_i$ is estimated by

$$\hat{D}Z_i'(Z_i\hat{D}Z_i' + \hat{\Sigma}_i)^{-1}(\mathbf{Y}_i - X_i\hat{\beta})$$

where the estimates are the REML estimates.

e) From part d) it follows that $\hat{\Sigma}_{\mathbf{Y}_i} = Z_i\hat{D}Z_i' + \hat{\Sigma}_i$. In this case $Z_i = (1,1,1)^T$ and $\hat{\Sigma}_i = \hat{\sigma}^2 I_3$. From the output $\hat{d} = 0.1346215r$ and $\hat{\sigma}^2 = 0.747435^2$. Hence, if $\mathbf{1}_3 = (1,1,1)^T$

$$\hat{\Sigma}_{\mathbf{Y}_i} = 0.0419192^2\mathbf{1}_3\mathbf{1}_3^T + 0.747435^2 I_3 =$$

$$\begin{pmatrix} 0.0419192^2 + 0.7505293^2 & 0.0419192^2 & 0.0419192^2 \\ 0.0419192^2 & 0.0419192^2 + 0.7505293^2 & 0.0419192^2 \\ 0.0419192^2 & 0.0419192^2 & 0.0419192^2 + 0.7505293^2 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.5650514 & 0.001757219 & 0.001757219 \\ 0.001757219 & 0.5650514 & 0.001757219 \\ 0.001757219 & 0.001757219 & 0.5650514 \end{pmatrix}.$$

SLUTT