UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

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

Day of examination: Monday December 1th 2014

Examination hours: 14.30 – 18.30

This problem set consists of 5 pages.

Appendices: None

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.

Solution proposal

Problem 1

1a

i) Show that if Y is a stochastic variable with a distribution belonging to the exponential family, then $\mathrm{E}(Y) = a'(\theta)$ and $\mathrm{Var}(Y) = \phi a''(\theta)$, where a' and a'' denote the first and second derivatives of a. [Hint: Start with calculating the first derivative of $f(y; \theta, \phi)$ with respect to θ.]

Proof for $\mathrm{E}[Y] = a'(\theta)$

First derivative of f: $f'(y; \theta, \phi) = \frac{y-a'(\theta)}{\phi} f(y; \theta, \phi)$

Integral of left side:

$$\int f'(y; \theta, \phi) dy = \frac{\partial}{\partial \theta} \int f(y; \theta, \phi) dy = \frac{\partial}{\partial \theta}(1) = 0$$

Integral of right side:

$$\frac{1}{\phi}(\int y f(y; \theta, \phi) dy - a'(\theta) \int f(y; \theta, \phi) dy) = \frac{\mathrm{E}[Y] - a'(\theta)}{\phi},$$

which gives $\mathrm{E}[Y] = a'(\theta)$.

We have here assumed that differentiation and integration can be interchange.

Proof for $\mathrm{Var}(Y) = \phi a''(\theta)$:

Second derivative: $f''(y; \theta, \phi) = \left[\left(\frac{y-a'(\theta)}{\phi} \right)^2 - \frac{a''(\theta)}{\phi} \right] f(y; \theta, \phi)$

(Continued on page 2.)

Integral of left side:

$$\int f''(y;\theta,\phi)dy = \frac{\partial^2}{\partial\theta^2}\int f(y;\theta,\phi)dy = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

Integral of right side:

$$\int \left[\left(\frac{y-a'(\theta)}{\phi}\right)^2 - \frac{a''(\theta)}{\phi}\right] f(y;\theta,\phi)dy = \frac{\mathrm{Var}(Y)}{\phi^2} - \frac{a''(\theta)}{\phi}$$

which gives $\mathrm{Var}(Y) = \phi a''(\theta)$.

1b

i) Show that the Poisson distribution belongs to the exponential family.

The probability mass function can be written

$$f(y;\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda) = \frac{1}{y!}\exp(y\log(\lambda) - \lambda),$$

and it therefore belongs to the exponential family with

- $\theta = \log(\lambda)$
- $a(\theta) = \lambda = \exp(\theta)$
- $\phi = 1$
- $c(y;\phi = 1) = \frac{1}{y!}$

ii) Show that $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

$\mathrm{E}[Y] = a'(\theta) = \exp(\theta) = \lambda$

$\mathrm{Var}(Y) = \phi a''(\theta) = \exp(\theta) = \lambda$

1c

i) Give an interpretation of the parameter β_1 or some transformation of it.

If we first calculate $\mu = \mu(x_1, x_2) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and then increase the value first explanatory by one and calculate $\mu' = \mu(x_1 + 1, x_2) = \exp(\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2)$, then the ratio $\mu'/\mu = \exp(\beta_1)$.

So, the expectation of the response increases by a multiplicative factor $\exp(\beta_1)$ when x_1 is increased by one unit and x_2 is unchanged.

The expectation of a Poisson distributed variable is also the rate of a Poisson process over a given time interval, so $\exp(\beta_1)$ can also be called a rate-ratio and β_1 can be called a log-rate-ratio.

ii) Assume then that $\beta_0 = 1$, $\beta_1 = 2$ and β_2=3 and predict the response Y for $x_1 = 1$ and $x_2 = 1$ and then for $x_1 = 2$ and $x_2 = 1$.

$\widehat{y} = \widehat{\mu} = \exp(1 + 2\cdot 1 + 3\cdot 1) = \exp(6) = 403.4288$

$\widehat{y} = \widehat{\mu} = \exp(1 + 2\cdot 2 + 3\cdot 1) = \exp(8) = 2980.958$

(Continued on page 3.)

1d

i) Explain what over-dispersion means in Poisson regression.

If the Poisson assumption holds, then for an observation Y_i, $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i)$. Over-dispersion occurs if $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i)$.

ii) Explain why the results above show that the current count data are over-dispersed.

The phihat which is reported here is an estimate of the dispersion factor ϕ, given by $\widehat{\phi} = (1/(n-p+1))\sum_i (y_i - \widehat{\mu}_i)^2/\widehat{\mu}_i$, where n is the number of observations and $p+1$ is the number or parameters estimated. Since phihat is much larger than 1, the data are over-dispersed.

iii) Discuss shortly two different possibilities for performing a more correct analysis than that given above.

Possibility 1 - Quasi-likelihood: Specify only a structure on the expectation and another on the variance, but do not assume a specific distribution. For instance, assume $\mathrm{Var}(Y_i) = \phi\mu_i$. The quasi-likelihood estimates are then exactly the same as the Poisson estimates, but their standard errors are adjusted.

Possibility 1 - Negative binomial distribution: Assume that the response follows a negative binomial distribution, which is a distribution for count data that allows over-dispersion. The variance structure is then $\mathrm{Var}(Y_i) = \mu_i + \theta\mu_i^2$, where θ is a positive parameter that controls the degree of over-dispersion.

Problem 2

2a

i) Give an interpretation of a regression coefficient β, or a transformation of it, in binary regression with logit link function.

Odds is defined as $p/(1-p)$, so $g(p) = \log(p/(1-p))$ is the log-odds. Furthermore the ratio of two odds is called an odds-ratio. If we compute a probability p for some values of the explanatory variables and a probability p' for the same values of the explanatory variables except for the j-th variable, which is increased by one, then $g(p)-g(p') = \log(p/(1-p))-\log(p'/(1-p')) = \log[p/(1-p)]/[p'/(1-p')] = \beta_j$. Therefore β_j is the log-odds-ratio and $\exp(\beta)$ is the odds-ratio for one unit increase in the j-th explanatory variable.

ii) Give then a simpler interpretation of β which holds approximately for small probabilities.

When both the probabilities p and p' are small, the odds ratio $[p/(1-p)]/[p'/(1-p')]$ is approximately p/p', which we can call a relative risk. Then $\exp(\beta_j)$ is the relative risk for one unit increase in the j-th explanatory variable.

(Continued on page 4.)

2b

i) Define AIC.

AIC = -2 log likelihood + 2 p where p is the number of parameters in the model. It gives a balance between the fit to the data and the number of parameters in the model.

ii) Which one of the models above would you choose based on the given results? Why?

The model `m1.probit` with only x_1 and the probit link has the lowest AIC value, and could therefore be chosen as the preferred model. But the model `m1.logit` with only x_1 and the logit link has almost the same AIC value. I personally think that models with logit link are easier to interpret, therefore I choose the `m1.logit` model.

2c

i) Define the two terms sensitivity and specificity.

Sensitivity: Proportion of correct predictions when true $Y_i = 1$

Specificity: Proportion of correct predictions when true $Y_i = 0$

ii) Describe what a ROC (Receiver Operating Characteristics) curve is, and draw a plot with one curve for a model with good classification performance and another model which is no better than random classification.

Compute the sensitivity and the specificity for different threshold values γ between 0 and 1. Plot sensitivity on the y-axis and (1-specificity) on the x-axis. A curve for a good model lies in the upper left corner. A curve for random classification is a straight line.

Problem 3

3a

i) Discuss whether the random effect term b_i is an important part of the model compared to other parts of the model.

The estimated variance of b_i is 0.0061 and its standard deviation is 0.0818. On the original scale, the factor $\exp(b_i)$ is between $\exp(-1.96 \cdot 0.0818) = 0.85$ and $\exp(1.96 \cdot 0.0818) = 1.18$ for 95 % of the individuals, and this can be a seen as an important difference between individuals. However, x_1 does also account for individual variation, and $\text{Var}(\beta_1 x_1)$ is estimated to be $2.06202^2 \cdot 0.087 = 0.369$ which is much higher than the variance of b_i. Other terms in the model have also much higher variance than b_i. So, compared to the rest of the model, the random term b_i can be neglected.

3b

i) Use the information you have to suggest simplifications or improvements of the model.

(Continued on page 5.)

The p value for the Wald test for $H_0 : \beta_3 = 0$ is 0.404, so x_3 can be deleted from the model. This is confirmed by the scatter plot with $\log(y)$ vs. x_3 which shows no obvious relation between $\log(y)$ and x_3.

The scatter plot of $\log(y)$ vs. x_4 shows a clear non-linear relationship between the two. One possibility can be to divide x_4 into two variables around the value of about 0.5 and estimate a model with two linear pieces joined at 0.5. Another possibility could be to constrain the curve to be flat for $x_4 > 0.5$. A third probability could be to include a second order term of x_4, i.e. x_4^2.

END