TBC, May 2018
1.5 hour Honours/ 2 hours M.Sc.

EXAMINATION FOR THE DEGREES OF M.A., M.SCI. AND B.SC. (SCIENCE)

Statistics – Generalized Linear Models – Solutions

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

1. (a) For the negative binomial distribution we have

$$f(y;\theta) = \binom{y+r-1}{r-1}\theta^r(1-\theta)^y = \exp\left[\log\binom{y+r-1}{r-1} + r\log\theta + y\log(1-\theta)\right]$$

which is of the form

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right]$$

with $a(y) = y$, natural parameter $b(\theta) = \log(1-\theta)$, $c(\theta) = r\log\theta$ and $d(y) = \log\binom{y+r-1}{r-1}$. **[3 MARKS]**

(b) For Y_i independent from the negative binomial distribution with $\mu_i = E(Y_i)$, we can write the model equation as

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},\ i = 1, \ldots, n.$$

[2 MARKS]

END OF QUESTION PAPER.

In vector-matrix form

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

[2 MARKS]

One suitable choice of link function is $g(\mu) = \log \mu$ which ensures a positive μ. **[1 MARK]**

(c) In a generalized linear model for counts it is often assumed that the response follows a Poisson distribution, for which the variance equals the mean. Overdispersion occurs when this does not happen in practice, and instead we have a variance that is larger than the mean. **[2 MARKS]**

Overdispersion may occur in the model if important explanatory variables are omitted, if the Y_i are correlated, if the link function is misspecified or because of other data complexities. **[2 MARKS]**

The negative binomial distribution has variance that is larger than the mean, so it can be used to model overdispersed data. **[1 MARK]**

In the output from Model 1 we see a large residual deviance which may be a sign of overdispersion.

[**Discussion of the value of the dispersion parameter in the negative binomial model fit or any other correct answers also get full credit.**]

[1 MARK]

(d) The coefficient of `child` is negative, which means that groups with children are expected to catch fewer fish on average (all other things being equal).

[2 MARKS]

(e) The coefficient of `camper` is positive, so we would expect a group with a campervan to catch more fish on average (all other things being equal). **[2 MARKS]**

(f) The expected count for a group with `child`$= 2$ and `camper`$=1$ is $\exp(1.0727 - 1.3753 \times 2 + 0.9094) = 0.4637$. **[2 MARKS]**

(g) [**MSc only**] The histogram shows a large concentration of zero counts, which could be causing overdispersion in the data. Zero-inflated models could be fitted to deal with the excess zeros. **[2 MARKS]**

END OF QUESTION PAPER.

2. (a) Y_i independent Bernoulli(p_i) with

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i,\ i = 1, \ldots, n$$

where x_i is the emotional stability score of the ith employee, and p_i is the probability that the ith employee is able to perform the group task.

[5 MARKS]

(b) The null hypothesis is H_0: the model is a good fit to the data. **[2 MARKS]**
The Hosmer-Lemeshow test statistic is $X^2_{HL} = 4.8$ which is smaller than $\chi^2(8; 0.95) = 15.507$ so we do not reject H_0 and conclude that there is no evidence of lack of fit in the model. **[2 MARKS]**

(c) A Wald test of $H_0 : \beta_1 = 0$ can be performed by comparing the z-statistic for the stability coefficient to a standard normal distribution. $z = 2.402 > z_{0.975} = 1.96$ or equivalently the p-value for it is $0.0163 < 0.05$ indicating that we should reject H_0 and conclude that the effect of the stability score is significant at the 5% level.

[2 MARKS]

(d) Point estimate for the odds: $\exp(\hat{\beta}_1) = \exp(0.018920) = 1.0191$. **[1 MARK]**
Approximate 95% CI for β_1:

$$0.018920 \pm 1.96 \times 0.007877 = (0.00348108, 0.03435892)$$

Approximate 95% CI for the odds of being able to perform in the task group:

$$(\exp(0.00348108), \exp(0.03435892)) = (1.003487, 1.034956)$$

[3 MARKS]

Interpretation: For every unit increase in the stability the odds of being able to perform in the task group get multiplied by a factor of somewhere between 1.003487 and 1.034956 (point estimate 1.0191). **[2 MARKS]**

(e) The model equation can be written in terms of the probability of being able to perform in a task group as follows:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

so

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(-10.308925 + 0.018920 \times 500)}{1 + \exp(-10.308925 + 0.018920 \times 500)} = 0.2996584.$$

[3 MARKS]

END OF QUESTION PAPER.

(f) [**MSc only**] Solve the following equation for x:

$$\log\left(\frac{0.7}{1-0.7}\right) = -10.308925 + 0.018920x$$
$$\Rightarrow 0.8472979 = -10.308925 + 0.018920x$$
$$\Rightarrow x = 589.6524.$$

[3 MARKS]

(g) [**MSc only**] A model for ordinal responses such as the proportional odds model would be appropriate in this case. [**Any other correct answer gets full marks**]
[2 MARKS]

3. (a) The Y_i are independent Poisson random variables with $E(Y_i) = \mu_i = \beta_1 x_{i1} + \beta_2 x_{i2}$ for $i = 1, \ldots, 6$. **[3 MARKS]**

The canonical link function $g(\mu_i) = \log \mu_i$ would not be appropriate here as it does not describe the physical process generating the data. **[2 MARKS]**

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ should be non-negative, as they are estimates of the image densities. Similarly the fitted values $\hat{y}_i$ should be non-negative. **[3 MARKS]**

(b) Using equation (1) with the estimated coefficients from Analysis 1 we have $\hat{y}_1 = 11.867 \times 0.78 + 7.073 \times 0.16 = 10.38794$. **[2 MARKS]**

The estimates of β_1 and β_2 and the fitted values all satisfy the required non-negativity constraint.

[2 MARKS]

(c) We can test the hypothesis $H_0 : \beta_0 = 0$ by comparing deviances from the two models. Let D_0 be the deviance of the model without intercept and D_1 the deviance of the model with intercept. $D_0 - D_1 = 1.033 - 0.978 = 0.045 < \chi^2(1; 0.95) = 3.84$ so we do not reject H_0 at the 5% significance level. Hence we conclude that there is no evidence in favour of the model with intercept over the model without intercept.
[5 MARKS]

(d) Using raw residuals would be inappropriate because the variance of the residuals would not be constant but would depend on the mean. Deviance or Pearson residuals can be used instead. **[3 MARKS]**

(e) [**MSc only**] Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \sim Po(\mu_i)$ where μ_i is given by (1). Then the log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}; \mathbf{y}) &= \sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!) \\ &= \sum y_i \log(\beta_1 x_{i1} + \beta_2 x_{i2}) - \sum(\beta_1 x_{i1} + \beta_2 x_{i2}) - \sum \log(y_i!) \end{aligned}$$

END OF QUESTION PAPER.

[2 MARKS]

To find the maximum likelihood estimator we set the partial derivatives equal to zero:

$$\frac{\partial l}{\partial \beta_1} = -\sum x_{i1} + \sum \left(\frac{y_i x_{i1}}{\beta_1 x_{i1} + \beta_2 x_{i2}} \right) = 0$$
$$\frac{\partial l}{\partial \beta_2} = -\sum x_{i2} + \sum \left(\frac{y_i x_{i2}}{\beta_1 x_{i1} + \beta_2 x_{i2}} \right) = 0$$

[2 MARKS]

These likelihood equations do not have a closed form solution and are solved numerically using the method of scoring (iteratively reweighted least squares algorithm).

[1 MARK]

(f) For the full model each y_i has its own parameter μ_i. The maximum likelihood estimators of the μ_i are obtained by differentiating the log-likelihood with respect to each μ_i and setting equal to zero:

$$\frac{\partial l}{\partial \mu_i} = \frac{y_i}{\mu_i} - 1 = 0 \Rightarrow \hat{\mu}_i = y_i.$$

[2 MARKS]

The maximised log-likelihood for the full model then is

$$\sum y_i \log y_i - \sum y_i - \sum \log(y_i!).$$

For the model of interest the maximised log-likelihood is

$$\sum \hat{y}_i \log y_i - \sum \hat{y}_i - \sum \log(y_i!)$$

where $\hat{y}_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$.

Therefore the deviance for the model of interest is

$$\begin{aligned} D &= 2\left[\sum y_i \log y_i - \sum y_i - \sum \log(y_i!) - \sum \hat{y}_i \log y_i + \sum \hat{y}_i + \sum \log(y_i!)\right] \\ &= 2\left[\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum (y_i - \hat{y}_i)\right] \end{aligned}$$

[2 MARKS]

The deviance for Model (1) is $D = 1.033$. For a goodness of fit test we compare D to the 95th percentile of the $\chi^2(6-2)$ distribution. $D = 1.033 < \chi^2(4; 0.95) = 9.488$ so the model appears to be a good fit to the data. **[2 MARKS]**

However, the chi-squared approximation may not be very good if the fitted values are small, and here we have $\hat{y}_5$ and $\hat{y}_6$ fairly small (both less than 5). **[2 MARKS]**

Total: 80

END OF QUESTION PAPER.