TBC, May 2019
1.5 hour Honours/ 2 hours M.Sc.

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

Statistics – Generalised Linear Models – Solutions

1. (a) For a single observation y from a distribution with p.d.f. or p.m.f. of the form $f(y) = \exp[yb(\theta) + c(\theta) + d(y)]$, the log-likelihood is

$$l(\theta; y) = yb(\theta) + c(\theta) + d(y)$$

[1 MARK]

and the score function is

$$U(\theta) = \frac{dl}{d\theta} = yb'(\theta) + c'(\theta).$$

[1 MARK]

Using the property $E[U(\theta)] = 0$ we have that

$$E(Y)b'(\theta) + c'(\theta) = 0 \Rightarrow E(Y) = -\frac{c'(\theta)}{b'(\theta)}.$$

[2 MARKS]

(b) For the Poisson distribution the probability mass function can be written in exponential family form as

$$f(y; \theta) = \frac{\theta^y e^{-\theta}}{y!} = \exp(y \log\theta - \theta - \log y!) = \exp(yb(\theta) + c(\theta) + d(y)$$

CONTINUED OVERLEAF/

with $b(\theta) = \log\theta$, $c(\theta) = -\theta$ and $d(y) = -\log y!$.

[2 MARKS]

Since $c'(\theta) = -1$ and $b'(\theta) = 1/\theta$, we have $E(Y) = -\dfrac{c'(\theta)}{b'(\theta)} = \theta$.

[2 MARKS]

(c) (i) A suitable model for these data would be a loglinear model with the response following the Poisson distribution. Let Y_{ijk} be the frequency in the (ijk)th cell of the contingency table where $i = 1$ for high and $i = 2$ for low consumption of tomatoes, $j = 1$ for high and $j = 2$ for low consumption of broccoli and $k = 1$ for high and $k = 2$ for low consumption of carrots.

[2 MARKS]

If the Y_{ijk} are assumed to be independent Poisson random variables with parameters $E(Y_{ijk}) = \mu_{ijk}$, then their sum will follow a Poisson distribution with parameter $E(n) = \mu = \sum_i \sum_j \sum_k \mu_{ijk}$. Suppose that the total number of observations n is fixed by the design of the study so that $\sum_i \sum_j \sum_k Y_{ijk} = n$. Then the joint distribution of the Y_{ijk} given their sum n is a multinomial distribution with probabilities $\theta_{ijk} = \mu_{ijk}/\mu$, so θ_{ijk} can be thought of as the probability of an observation in the (ijk)th cell of the table.

[2 MARKS]

Since the expected value of Y_{ijk} is $E(Y_{ijk}) = n\theta_{ijk}$, taking the log link function gives

$$\begin{aligned}\log\mu_{ijk} &= \log n + \log\theta_{ijk}\\ &= \mu + \alpha_i + \beta_j + \gamma_j + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}\end{aligned}$$

where it would be of interest to explore which (if any) of the interaction terms are significant.

[2 MARKS]

(ii) Let Y_i be the number of serious infections at the ith hospital. The Y_i are assumed to be Poisson-distributed with $E(Y_i) = \mu_i = E_i\theta_i$ where the exposure factor, E_i, is equal to the average number of persons/day in the ith hospital, multiplied by the number of days that the hospital was in the study.

[3 MARKS]

The dependence on the covariates can be modelled as $\theta_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i_3})$ where $x_{i1} = 1$ for teaching and 0 for non-teaching, $x_{i2} = 1$ for public and 0 for private, and $x_{i3} = 1$ for religious and 0 for secular. The generalised linear model thus is

$$E(Y_i) = \mu_i = E_i \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i_3}); \qquad Y_i \sim Po(\mu_i).$$

CONTINUED OVERLEAF/

Equivalently, the model can be written as

$$\log(\mu_i) = \log(E_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

where $\log(E_i)$ is the offset term in the model.

[3 MARKS]

2. (a) Model `m1` includes an interaction term between $\log_{10}$(Dose) and preparation, but as this term is not significant (p-value of 0.549), there is no evidence that the effect of preparation depends on the dose. **[2 MARKS]**

(b) As the interaction term in model `m1` is not significant, we can choose the additive model `m0`. **[2 MARKS]**

(c) Regression equations corresponding to model `m0`:
Standard preparation: log odds of response $= -5.5531 + 5.2894 \log_{10}(\text{Dose})$

[2 MARKS]

Test preparation: log odds of response $= -5.5531 - 0.9290 + 5.2894 \log_{10}(\text{Dose})$

$$= -6.4821 + 5.2894 \log_{10}(\text{Dose})$$

[2 MARKS]

(d) (i) The median effective dose is the value which sets the probability of response equal to 0.5 or, equivalently, the log odds of response equal to 0. So, for the standard preparation solving
$\log_{10}$(Median Effective Dose)=0 gives

$$\begin{aligned} -5.5531 + 5.2894 \log_{10}(\text{Median Effective Dose}) &= 0 \\ \Rightarrow \log_{10}(\text{Median Effective Dose}) &= \frac{5.5531}{5.2894} = 1.049854 \\ \Rightarrow \text{Median Effective Dose} &= 10^{1.049854} = 11.2. \end{aligned}$$

[2 MARKS]

(ii) Similarly for the test preparation $\log_{10}(\text{Median Effective Dose}) = \frac{6.4821}{5.2894} = 1.225489 \Rightarrow \text{Median Effective Dose} = 10^{1.225489} = 16.8$.

[2 MARKS]

(e) Estimated potency= 11.2/16.8 = 0.6666667. **[2 MARKS]**

CONTINUED OVERLEAF/

(f) If the actual dose is doubled, the log odds of response will increase by $5.2894 \times \log_{10}(2) = 1.592268$ and so the odds of response will get multiplied by $\exp(1.592268) = 4.9$. Thus the required odds ratio is 4.9. **[3 MARKS]**

(g) Based on the residual deviance of 8.7912 on 11 degrees of freedom there is no evidence of lack of fit for model `m0` since $8.7912 < \chi^2(11; 0.95) = 19.7$. We have to be a bit cautious about using the chi-squared approximation as there are a few small fitted values (obs 1, 2 and 10).

Note: credit will be given for any sensible comments on the fit of the model based on residuals, X^2 statistic or similar.

[3 MARKS]

3. (a) The offset accounts for the different amounts of exposure for the doctors. The number of visits and hence the number of patients seen during the study period varies from doctor to doctor and needs to be taken into account before any comparisons can be made.

[2 MARKS]

(b) The coefficient of residency is negative, indicating that doctors who had completed residency training had a lower complaint rate than their colleagues who had not completed residency training. The rate ratio corresponding to the residency coefficient is $\exp(-0.3041) = 0.74$, so the rate of complaints for doctors with residency training is 0.74 times that of doctors without residency training.

[2 MARKS]

The effect of residency is not significant at the 5% level, as the p-value is 0.078. (Answers commenting on the effect being marginally significant will also be accepted.)

[2 MARKS]

(c) "Fisher Scoring iterations" gives the number of iterations it took for the Iteratively Reweighted Least Squares algorithm to converge. This is the algorithm that solves the likelihood equations numerically in order to obtain the MLE $\hat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$.

[2 MARKS]

(d) The reduction in deviance from the null model to that with H only is $63.435 - 57.347 = 6.088$ which is significant when compared with $\chi^2_{0.95}(1) = 3.84$. The deviance drop going from H to H+R is $57.347 - 57.131 = 0.216$ which is not significant when compared with $\chi^2_{0.95}(1) = 3.84$.

Comparing H to H+P+R, the deviance drop is $57.347 - 57.341 = 0.006$ which is not significant when compared with a $\chi^2_{0.95}(1) = 5.99$.

CONTINUED OVERLEAF/

From the models with main effects given in the table, the model with just H appears to be the best.

[2 MARKS]

Comparing H to H+P+R+H*P the drop in deviance is $57.347 - 53.789 = 3.558$ which is not significant when compared to a $\chi^2_{0.95}(3) = 7.81$. Similarly H+P+R+H*R is not preferred to the model with just H as $57.347 - 50.182 = 7.162 < 7.81$.

[2 MARKS]

The next comparison is between H and H+P+R+H*P+H*R . The drop in deviance is $57.347 - 44.747 = 12.6$ which is greater than $\chi^2_{0.95}(4) = 9.49$ so the model with the two-way interaction terms is preferred.

Finally adding a third two-way interaction to the model (P*R) only changes the deviance by a small amount $(44.747 - 44.405)$ so there is no need to add the term P*R. The model H+P+R+H*P+H*R is chosen as the final model.

[2 MARKS]

(e) In GLMs for binomial and Poisson responses, the mean and variance of the response are related to each other. For Poisson regression the variance is equal to the mean, so when we specify a model for the relationship between the mean and explanatory variables of interest, we are also imposing the same model on the variance. Overdispersion occurs when the observed counts exhibit more variability than the model allows for.

[2 MARKS]

Possible causes of overdispersion in a Poisson model for counts:

- Omission of important explanatory variables from the model
- Misspecification of the model in some other way, *e.g.* the link function is wrong
- Excess zeros in the data
- Lack of independence, e.g. due to clustered responses
- More complex structure of the model

Any two of the above, or any other sensible answer(s) will receive full credit.

[2 MARKS]

Ways in which we can deal with overdispersion:

- Add more explanatory variables to the model, if data are available
- Use a quasi-Poisson model to adjust the standard errors of the estimates
- Use a negative binomial instead of a Poisson distribution for the response

Any two of the above, or any other sensible answer(s) will receive full credit.

[2 MARKS]

CONTINUED OVERLEAF/

4. (a) The likelihood for the null model is

$$L(p; y) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

where $p = P(Y_i = 1)$ and $1 - p = P(Y_i = 0)$.
The log-likelihood is

$$\begin{aligned} l(p) &= \sum_{i=1}^{n} y_i \log p + \sum_{i=1}^{n}(1-y_i)\log(1-p) \\ &= \sum_{i=1}^{n} y_i \log p + n\log(1-p) - \sum_{i=1}^{n} y_i \log(1-p) \end{aligned}$$

[2 MARKS]

Differentiate with respect to p to get the likelihood equation:

$$\frac{\sum_{i=1}^{n} y_i}{\hat{p}} - \frac{n - \sum_{i=1}^{n} y_i}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y}{n} \text{ where } y =_{i=1}^{n} y_i.$$

[1 MARK]

Substituting $\dfrac{y}{n}$ into the log-likelihood we have:

$$\begin{aligned} &\sum_{i=1}^{n} y_i \log\frac{y}{n} + n\log(1-\frac{y}{n}) - \sum_{i=1}^{n} y_i \log(1-\frac{y}{n}) \\ &= y\log y - y\log n + n\log(n-y) - n\log n - y\log(n-y) + y\log n \\ &= y\log y + (n-y)\log(n-y) - n\log n. \end{aligned}$$

[2 MARKS]

Here we have $y = 20$ and $n = 53$ so the maximised null log-likelihood is

$$\begin{aligned} &y\log y + (n-y)\log(n-y) - n\log n \\ &= 20\log(20) + 33\log(33) - 53\log(53) \\ &= -35.126. \end{aligned}$$

The null deviance is calculated as twice the maximised log-likelihood for the saturated model minus that of the null model, so it is equal to $2[0-(-35.126)] = 70.252$ as required.

[2 MARKS]

CONTINUED OVERLEAF/

(b) The lowest predicted probability of nodal involvement for any future patient will be obtained if we set `age=1`, and all other explanatory variables=0. This is because age has a negative coefficient while all the other predictors have positive coefficients.

[2 MARKS]

Then the predicted probability can be calculated as:

$$\frac{\exp(-3.079 - 0.292)}{1 + \exp(-3.079 - 0.292)} = 0.033.$$

[2 MARKS]

(c) Possible next steps:

- Dropping age as it is not significant.
- Variable selection more generally, deciding which variables to keep using Wald or generalised likelihood ratio tests.
- Considering models which include interaction terms in addition to the above main effects.
- If the variables are also available in non-binary format (*e.g.* age, acid), explore the possibility of using continuous versions of these explanatory variables.

Any of the above or any other sensible answer will receive full credit.

[3 MARKS]

(d) If $\hat{p}_i > c$, assign the subject to the class with nodal involvement present. If $\hat{p}_i \leq c$, assign the subject to the class without nodal involvement. Here the threshold c can be taken as 0.5 or some other value.

[2 MARKS]

The predictive performance of the model can be assessed in terms of sensitivity (probability of correctly predicting nodal involvement when it is present), specificity (probability of correctly predicting no nodal involvement when it is absent) and overall accuracy (number of correct classifications of either type over total number of subjects).

[2 MARKS]

A Receiver Operating Characteristic (ROC) curve can be used to explore the effect of changing the threshold c on the sensitivity and specificity. This would allow selection of a value of c that maximises the accuracy of the classification procedure.

[2 MARKS]

Total: 80

END OF QUESTION PAPER.