University of Glasgow

20 May 2019
1.5 hour Honours/ 2 hours M.Sc.

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

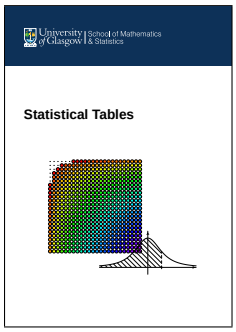
Statistics – Generalised Linear Models – Solutions

This paper consists of 8 pages and contains 4 questions.
Candidates should attempt all questions.

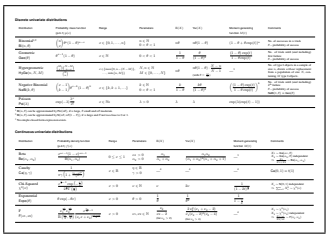
Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Question 4	20 marks
Total	80 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

CONTINUED OVERLEAF/

1. (a) Generalised linear models extend the normal linear model by

- allowing the response to follow distributions other than the normal distribution;

[2 MARKS]

- setting a more general function g of the mean equal to the linear predictor, so that instead of $\mu = \mathbf{x}_i^\intercal \beta$ we have $g(\mu) = \mathbf{x}_i^\intercal \beta$. **[2 MARKS]**

(b) For a single observation y from a distribution with p.d.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 $E(U) = 0$ we have that

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

[1 MARK]

(c) The p.d.f. of the exponential distribution is

$$f(y; \theta) = \theta \exp(-\theta y) = \exp(\log \theta - \theta y)$$

[1 MARK]

This is in the form given in part (b) with $a(y) = y$,

$$b(\theta) = -\theta,$$

[1 MARK]

and

$$c(\theta) = \log \theta.$$

[1 MARK]

(d)

$$\begin{aligned} b(\theta) = -\theta \Rightarrow b'(\theta) &= -1 \\ c(\theta) = \log \theta \Rightarrow c'(\theta) &= \frac{1}{\theta} \\ E(Y) = -\frac{c'(\theta)}{b'(\theta)} &= -\frac{1/\theta}{-1} = \frac{1}{\theta}. \end{aligned}$$

[3 MARKS]

CONTINUED OVERLEAF/

Canonical link function: we need $b(\theta) \propto \mathbf{x}_i^\intercal \beta$, the linear component. Since $\mu = E(Y) = 1/\theta$ and $b(\theta) = -\theta$, we take $\mathbf{x}_i^\intercal \beta$ to equal the inverse link, $g(\mu) = \dfrac{1}{\mu}$.

[1 MARK]

(e) We have Y_i independent exponential random variables with $\mu_i = E(Y_i) = \dfrac{1}{\theta_i}$ and $\dfrac{1}{\mu_i} = \beta_0 + \beta_1 x_i$. The log-likelihood is

$$\begin{aligned} l(\theta_1, \ldots, \theta_n; \mathbf{y}) &= \sum_{i=1}^{n} \left[\log \theta_i - \theta_i y_i\right] = \sum_{i=1}^{n} \left[\log\left(\frac{1}{\mu_i}\right) - \frac{y_i}{\mu_i}\right] \\ &= \sum_{i=1}^{n} \left[\log(\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 x_i) y_i\right]. \end{aligned}$$

[1 MARK]

For the score vector we need the first partial derivatives:

$$\frac{\partial l(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} \left[\frac{1}{\beta_0 + \beta_1 x_i} - y_i\right]$$

and

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} \left[\frac{x_i}{\beta_0 + \beta_1 x_i} - x_i y_i\right].$$

[1 MARK]

For the information matrix we need the second derivatives:

$$\frac{\partial^2 l}{\partial \beta_0^2} = \frac{\partial}{\partial \beta_0}\left\{\sum_{i=1}^{n} \frac{1}{\beta_0 + \beta_1 x_i} + y_i\right\} = -\sum_{i=1}^{n} \left[\frac{1}{(\beta_0 + \beta_1 x_i)^2}\right];$$

$$\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n} \frac{x_i}{(\beta_0 + \beta_1 x_i)^2};$$

and

$$\frac{\partial^2 l}{\partial \beta_1^2} = -\sum_{i=1}^{n} \frac{x_i^2}{(\beta_0 + \beta_1 x_i)^2}.$$

[1 MARK]

This gives the information matrix

$$\mathcal{I} = \begin{pmatrix} \sum_{i=1}^{n} \frac{1}{(\beta_0+\beta_1 x_i)^2} & \sum_{i=1}^{n} \frac{x_i}{(\beta_0+\beta_1 x_i)^2} \\ \sum_{i=1}^{n} \frac{x_i}{(\beta_0+\beta_1 x_i)^2} & \sum_{i=1}^{n} \frac{x_i^2}{(\beta_0+\beta_1 x_i)^2} \end{pmatrix}.$$

CONTINUED OVERLEAF/

[1 MARK]

The score vector and information matrix evaluated at the current value of $\hat{\beta}$ play a role in the updating equation

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1}[\mathbf{U}^{(m-1)}]$$

which is used in the method of scoring/ iteratively reweighted least squares algorithm to solve for the MLE numerically.

[2 MARKS]

2. (a) It would be reasonable to expect that there would be more fatalities in traffic accidents in states where people drive more, so to be able to compare states with each other, it is important to take into account the total miles travelled in each state. By including the term log(milestot) as an offset in the model, the quantity modelled now is the fatality rate per million miles travelled, which makes comparisons possible between states with different amounts of driving.

[2 MARKS]

(b) Given that the `year` coefficients are negative and mostly increasing in absolute value, there appears to be a decline in the fatality rate over time. **[1 MARK]**

The rate ratio comparing the expected fatality rate in 1988 to that from 1982 (and controlling state and beer tax) is obtained as $\exp(-0.196689) = 0.82$. There was an 18% reduction in the fatality rate from 1982 to 1988. **[2 MARKS]**

(c) The rate ratio for `beertax` is $\exp(-0.292140) = 0.747$ with an approximate 95% CI of $\exp(-0.292140 \pm 1.96 \times 0.037654) = (0.694, 0.804)$. For each increase of one percentage point in beer tax, the fatality rate is multiplied by a factor of 0.747 (confidence interval of 0.694,0.804), i.e. a 25% reduction. **[3 MARKS]**

(d) A Poisson GLM assumes that the variance of the responses, Y_i, is equal to their mean, μ_i. In practice, there may be more variability in the data than the model allows for, so that $\text{Var}(Y_i) > \mu_i$. This is called overdispersion. **[2 MARKS]**

In the above model overdispersion may arise because other relevant explanatory variables may have been omitted from the model. **[2 MARKS]**

[**Note:** Any other reasonable answer will receive credit here.]

(e)
- Plot $(y_i - \hat{\mu}_i)^2$ against $\hat{\mu}_i$ for a visual check of whether the majority of points lie above the line of equality. **[2 MARKS]**
- Estimate the dispersion parameter ϕ by dividing $X^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$ by $n - p = 281$. A large estimate of ϕ is an indication of overdispersion. **[2 MARKS]**

(f) Any two of the following or any other reasonable answer gets full marks.

CONTINUED OVERLEAF/

- Add more explanatory variables to the model.
- Fit a negative binomial model to adjust for overdispersion.
- Fit a model that accounts for correlated observations from the same state.

[4 MARKS]

3. (a) Consider $\Pr(Z < x_i)$ in the case where $Z \sim N(-\beta_0/\beta_1; 1/\beta_1^2)$ and $\beta_1 > 0$:

$$\Pr(Z < x_i) = \Pr\left(\frac{Z - (-\beta_0/\beta_1)}{1/\beta_1} < \frac{x_i - (-\beta_0/\beta_1)}{1/\beta_1}\right) = \Phi\left(\frac{x_i - (-\beta_0/\beta_1)}{1/\beta_1}\right) = \Phi(\beta_1 x_i + \beta_0)$$

as required. In the case $Z \sim N(-\beta_0/\beta_1; 1/\beta^2)$ and $\beta_1 < 0$ we have

$$\begin{aligned}\Pr(Z > x_i) &= \Pr[Z - (-\beta_0/\beta_1) > x_i - (-\beta_0/\beta_1)] \\ &= \Pr\left(\frac{Z - (-\beta_0/\beta_1)}{1/\beta_1} < \frac{x_i - (-\beta_0/\beta_1)}{1/\beta_1}\right) \\ &= \Phi\left(\frac{x_i - (-\beta_0/\beta_1)}{1/\beta_1}\right) = \Phi(\beta_1 x_i + \beta_0).\end{aligned}$$

[4 MARKS]

Using $p_i = \Phi^{-1}(\beta_0 + \beta_1 x_i)$ ensures $0 \leq p_i \leq 1$. This may not be true for $p_i = \beta_0 + \beta_1 x_i$. **[2 MARKS]**

(b) The fitted probability of success for group 2 can be obtained using

$$\hat{p}_2 = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_2)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_2)} = \frac{\exp(1.4087 - 0.5855 \times 1.1)}{1 + \exp(1.4087 - 0.5855 \times 1.1)} = 0.682$$

[2 MARKS]

The fitted number of successes is

$$\hat{y_2} = n_2 \hat{p}_2 = 9 \times 0.682 = 6.14.$$

[1 MARK]

(c) Interpretation of the estimated coefficient $\hat{\beta}_1$: the odds of a success get multiplied by $\exp(-0.5855) = 0.56$ for every unit increase in dose (so the higher the dose the lower the odds of success). **[2 MARKS]**

(d) Considering raw residuals to check the model fit would be inappropriate as $\text{Var}(Y_i) = n_i p_i (1 - p_i)$ which is not constant for all observations. **[1 MARK]**
Instead one could look at deviance or Pearson residuals. **[1 MARK]**

CONTINUED OVERLEAF/

(e) i. To test the goodness of fit of the model with $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ we compare the deviance from this model to the 95th percentile of the $\chi^2(4)$ distribution. Deviance $= 1.52 < 9.48$ so there is no evidence of lack of fit. **[3 MARKS]**

This relies on the asymptotic distribution of the deviance which requires reasonably large expected frequencies (fitted values) for all dose groups. **[1 MARK]**

ii. Hypothesis test of $\beta_1 = 0$: Take the difference in deviances from the model with dose included as an explanatory variable and the null model: $D_0 - D_1 = 7.15 - 1.52 = 5.63 > \chi^2_{0.95}(5-4) = 3.84$. We therefore reject the null hypothesis and conclude that dose is significant. **[3 MARKS]**

4. [MSc only]

(a) The likelihood 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 and solve for the MLE:

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

[2 MARKS]

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

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

CONTINUED OVERLEAF/

[2 MARKS]

(b) Maximised log-likelihood under the null model: $= y \log y + (n-y)\log(n-y) - n \log n = 300\log(300) + 700 \times \log(700) - 1000\log(1000) = -610.86$

Null deviance $= -2 \times (-610.86) = 1221.729$

[2 MARKS]

(c) The table including the number of parameters in each model is given below:

Variables in model	Deviance	Number of parameters in model
Null	1221.73	1
C	1199.06	5
A	1090.39	2
D	1177.11	2
C + A	1070.47	6
C + D	1051.90	6
A + D	1176.55	3
C + A + D	1051.19	7

Using backward elimination, we would reject the model $A + D$ in favour of the model with all three variables since the difference in deviance between the two models is $1176.55 - 1051.19 = 125.36$ which is significant when compared with $\chi^2_{0.95}(7-3) = 9.48$. Similarly we would reject model $C + A$ in favour of $C + A + D$ as the difference in deviance is $1070.47 - 1051.19 = 19.28$ which is significant when compared with $\chi^2_{0.95}(7-6) = 3.84$. However, model $C + D$ is not rejected in favour of the model with all three variables, since the difference in deviance between the two models is $1051.90 - 1051.19 = 0.71$ which is not significant when compared with $\chi^2_{0.95}(7-6) = 3.84$.

Dropping D from model $C + D$ increases the deviance significantly as $1199.06 - 1051.90 = 147.16$ is significant when compared with $\chi^2_{0.95}(6-4) = 5.99$. Similarly dropping C from model $C + D$ increases the deviance significantly as $1177.11 - 1051.90 = 125.21$ is significant when compared with $\chi^2_{0.95}(6-2) = 9.48$.

We therefore conclude that the simplest model that best describes the data is the one with $C + D$. **[5 MARKS]**

(d) Let p_i be the probability of loan approval for the ith applicant. The selected model is of the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 C_{1i} + \beta_2 C_{2i} + \beta_3 C_{3i} + \beta_4 C_{4i} + \beta_5 D_i$$

where C_1, C_2, C_3 and C_4 are dummy variables for levels 1-4 of factor C, and D_i is the loan duration for the ith applicant. **[2 MARKS]**

CONTINUED OVERLEAF/

By obtaining $\hat{p}$ from the above model based on the applicant's borrower characteristics, the bank can apply a rule such as "if $\hat{p} > c$ approve the loan application". **[2 MARKS]**

(e) Using the additional 1000 observations as a test set, the predictive performance of the model can be assessed using the following measures:

- overall accuracy: the proportion of correct predictions in the test dataset, the higher the better;
- false positive error rate: the proportion of time a negative outcome was incorrectly predicted as positive by the model for the test data, the lower the better;
- false negative error rate: the proportion of time a positive outcome was incorrectly predicted as negative by the model for the test data, the lower the better.

Note: Any other appropriate answer will receive credit. In particular, reference may be made to Receiver Operating Characteristic (ROC) curves and the area under the curve (AUC). **[3 MARKS]**

END OF QUESTION PAPER.