



The  
University  
Of  
Sheffield.

**MAS61001**

**SCHOOL OF MATHEMATICS AND STATISTICS**

**Spring Semester 2020–2021**

**Generalised Linear Models**

**2.5 Hours**

*This is an open book exam.*

*Answer both questions.*

*You can work on the exam during the 24 hour period starting from 10am (BST), and you must submit your work within 2.5 hours of accessing the exam paper or by the end of the 24 hour period (whichever is earlier).*

***Late submission will not be considered without extenuating circumstances.** Calculations should be performed by hand. A university-approved calculator may be used. The use of any other calculational device, software or service is not permitted. To gain full marks, you will need to show your working. By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged, and that no unfair means have been used.*

- 1 (a) Consider that observations  $y_i$  are generated by the exponential family of distributions

$$f(y_i) = \exp \left[ w_i \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right], \quad (1)$$

where  $w_i$  are the weights,  $\phi$  is the scale parameter,  $\theta_i$  is the natural parameter,  $b(\theta_i)$  is a known function of  $\theta_i$  and  $c(y_i, \phi)$  is a known function of  $y_i, \phi$ . Assume that  $y_i$  is independent of  $y_j$ , for  $i \neq j$ , and that the mean  $\mu_i = E(y_i)$  is mapped to the linear predictor  $\eta_i = x_i^T \beta$  via the link function  $g(\cdot)$ , so that  $g(\mu_i) = \eta_i$ .

- (i) Based on a set of data  $y = (y_1, y_2, \dots, y_n)$ , write down the log-likelihood function  $\ell(\mu; y)$  of  $\mu = [\mu_1, \mu_2, \dots, \mu_n]^T$ . **(1 mark)**
- (ii) Using the canonical link, calculate the partial derivatives of  $\ell(\mu; y)$ , with respect to  $\beta_k$ , for  $k = 1, 2, \dots, p$ . **(6 marks)**
- (iii) Using part (ii) show that the MLE  $\hat{\mu}$  of  $\mu$  satisfies the matrix equality

$$X^T W y = X^T W \hat{\mu},$$

where  $X$  is the design matrix and  $W$  is the weight matrix, defined below

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}. \quad (3 \text{ marks})$$

- (b) Consider the discrete random variable  $Y_i = y_i$  (taking values  $1, 2, 3, \dots$ ) generated by the following distribution

$$f(y_i) = \Pr[Y_i = y_i] = (1 - \pi_i)^{y_i - 1} \pi_i, \quad y = 1, 2, \dots, \quad 0 \leq \pi_i \leq 1,$$

where  $i = 1, 2, \dots, n$ .

- (i) Write  $f(y_i)$  in exponential form (1) and hence determine  $\theta_i, b(\theta_i), c(y_i, \phi), \phi$  and  $w_i$ . **(2 marks)**
- (ii) Use part (i) to calculate the mean  $E(Y_i)$  and the variance  $\text{Var}(Y_i)$  of  $Y_i$ . **(3 marks)**
- (iii) Show that the canonical link of  $f(y_i)$  is

$$g(\mu_i) = \log \frac{\mu_i - 1}{\mu_i},$$

where  $\mu_i = E(Y_i)$ . **(3 marks)**

- (iv) Show that the  $X^2$  statistic is

$$X^2 = \sum_{i=1}^n \frac{(y_i \hat{\pi}_i - 1)^2}{1 - \hat{\pi}_i},$$

where  $\hat{\pi}_i$  is the estimate of  $\pi_i$ . **(2 marks)**

- 2 A study is conducted on 8525 patients to investigate the effect of a particular treatment on lung cancer remission. The data recorded, summarised in the table below, are remission (signs and symptoms of cancer reduced, binary), C-reactive protein (CRP) as a diagnostic of cancer progression, duration of treatment (in months), cancer stage (coded as stage I, II, III, IV) and the ID of the doctor administering/overseeing the treatment.

Variable	type	description
remission	binary	1=symptoms free and 0=symptoms still exist
CRP	continuous	the value of CRP (milligram per litre or mg/L)
CancerStage	factor	Levels are stage I, II, III, IV
LengthofStay	discrete	number of months patient received treatment
DID	factor	each patient's doctor ID (407 different doctors)

- (a) A first analysis involves fitting a generalized linear model with `remission` as the response variable. Part of the R output of this analysis is given below:

Call:

```
glm(formula = remission ~ CRP + CancerStage + LengthofStay +
     DID, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )							
(Intercept)	-1.747e+01	1.211e+03	-0.014	0.988488							
CRP	-2.135e-02	1.049e-02	-2.035	0.041890	*						
CancerStageII	-4.175e-01	7.785e-02	-5.363	8.19e-08	***						
CancerStageIII	-1.044e+00	1.008e-01	-10.362	< 2e-16	***						
CancerStageIV	-2.403e+00	1.626e-01	-14.778	< 2e-16	***						
LengthofStay	-1.268e-01	3.452e-02	-3.673	0.000239	***						
DID2	1.820e+01	1.211e+03	0.015	0.988006							
DID3	1.994e+01	1.211e+03	0.016	0.986863							
.....											
DID406	-2.617e-01	3.924e+03	0.000	0.999947							
DID407	2.076e+01	1.211e+03	0.017	0.986320							
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 10352.6 on 8524 degrees of freedom
Residual deviance: 6133.7 on 8113 degrees of freedom
> qchisq(0.95,8524)
[1] 8739.896
> qchisq(0.95,8113)
[1] 8323.654
```

- (i) Explain why the degrees of freedom of the null model is 8524 and why the degrees of freedom of the current model is 8113.

(2 marks)

2 (continued)

- (ii) Formally compare the null and current models and evaluate the goodness of fit of the preferred one. *(4 marks)*
- (iii) Suggest improvements that you can make to this model (list at least two) and justify your answer. *(2 marks)*

(b) A second analysis of data in R involves the following output.

```
glmer(remission ~ CRP + CancerStage + LengthofStay +
      (1 | DID), family = binomial)
```

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
Family: binomial ( logit )
Formula: remission ~ CRP + CancerStage + LengthofStay + (1 | DID)
```

AIC	BIC	logLik	deviance	df.resid
7437.1	7486.5	-3711.6	7423.1	8518

Random effects:

Groups Name	Variance	Std.Dev.
DID (Intercept)	4.282	2.069

Number of obs: 8525, groups: DID, 407

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.167	0.209	-0.804	0.421229
CRP	-0.021	0.010	-2.108	0.035012 *
CancerStageII	-0.414	0.076	-5.479	4.27e-08 ***
CancerStageIII	-1.006	0.098	-10.253	< 2e-16 ***
CancerStageIV	-2.325	0.158	-14.741	< 2e-16 ***
LengthofStay	-0.119	0.033	-3.537	0.000405 ***

- (i) Write down the algebraic expression of this model ( $y_i = \dots$ ). *(3 marks)*
- (ii) Provide some evidence to support the claim that this model performs better than the model of part (a). *(2 marks)*
- (iii) Give a 95% confidence interval for the odds ratio of remission (= 1), for a 2-month increase of LengthofStay adjusting for the variables CRP and CancerStage. *(4 marks)*
- (iv) A patient on the cancer stage III has CRP value 7.2 mg/L and has received treatment for 6 months. Calculate the probability of the cancer be reduced (remission=1) for this patient. *(3 marks)*

**End of Question Paper**