UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Problem Sheet 5 **Semester 1, 2022–2023**

Work on Questions 1 and 2 in the workshop.

1. Let $Y_1, \ldots, Y_n$ be independent, with Y_i having a Poisson distribution and mean

$$\lambda_i = \theta_1 + \theta_2 x_i,$$

for fixed x_i, $i = 1, \ldots, n$. The following discrete data on textile faults, which has $n = 32$ measurements of cloth length (x) and the number of faults (y), is modelled by such a Poisson regression model.

x:	551	651	832	375	715	868	271	630	491	372	645	441	895	458	642	492
y:	6	4	17	9	14	8	5	7	7	7	6	8	28	4	10	4
x:	543	842	905	542	522	122	657	170	738	371	735	749	495	716	952	417
y:	8	9	23	9	6	1	9	4	9	14	17	10	7	3	9	2

   (a) Show that this problem can be treated as a generalized linear model (GLM).

   (b) Fit the model in R.

   (c) Calculate, from the fitted values given by R, the Pearson residual corresponding to the first point $(551, 6)$.

   (d) To perform a diagnostic check on the adequacy of the above model, plot the Pearson residuals against x in R.

   Comment on the results.

2. (*Question 1 continued:*) Fit the Poisson regression model with $\lambda_i = \theta_1$ by hand.

   Calculate the deviance for this model and also the model with $\lambda_i = \theta_1 + \theta_2 x_i$. Hence, use *analysis of deviance* to test the significance of the parameter θ_2, given that the model with mean $\lambda_i = \theta_1 + \theta_2 x_i$ is adequate.

3. The gamma distribution with probability density function given by

$$f(y;\theta) = \frac{y^{\phi-1}\theta^{\phi}e^{-y\theta}}{\Gamma(\phi)}, \quad (\phi \text{ known}),$$

has mean $\frac{\phi}{\theta}$ and variance $\frac{\phi}{\theta^2}$. The fitted values for a GLM with gamma responses and link function g are given by

$$\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T \widehat{\beta}),$$

$i = 1, \ldots, n$, and the responses are $y_1, \ldots, y_n$.

Obtain expressions, in terms of response values and fitted values, for

   (a) The deviance.

   (b) The Pearson residuals.

4. The data in the following table give the number of positive responses of a binary $0/1$ variable, out of 50, for 5 different values of a continuous explanatory variable x.

x	Number of trials (N)	Number of responses (R)
1	50	0
2	50	12
3	50	25
4	50	45
5	50	49

The probability of response, θ_i, for a given x_i, is modelled by a relationship of the form

$$\theta_i = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}.$$

Show that this model can be expressed in the form

$$\text{logit}(\theta_i) = \alpha+\beta x_i, \quad \text{where} \quad \text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right).$$

This is a GLM with binomial family and logit link.

(a) Estimate the unknown parameters α and β of the model using ordinary least squares, taking $\text{logit}(\widetilde{\theta})$, where $\widetilde{\theta} = (R+\frac{1}{2})/N$, as the dependent variable.

(b) Fit the GLM model in R.

(c) Calculate the fitted values for the model with $\text{logit}(\theta_i) = \alpha$ (with and/or without use of R).

(d) Calculate the deviances for the two models by hand.

Use analysis of deviance to

(i) Test for evidence of nonlinearity, i.e., test the goodness-of-fit of the full model with $\text{logit}(\theta_i) = \alpha+\beta x_i$.

(ii) Assuming that the full model is adequate, test $H_0 : \beta = 0$.

FURTHER EXERCISES (GLMs):

5. (a) Let p_{ij} denote the probability that an observation falls into cell (i, j) of a two-way contingency table containing r rows and c columns. A total of $x_{\bullet\bullet}$ observations are taken and x_{ij} denotes the number that fall into cell (i, j).

   Obtain an expression for the maximum likelihood estimate of p_{ij}

   (i) under the assumption that the rows and columns are independent;

   (ii) without the restriction of independence of rows and columns.

   (b) Using your answers to (a), show that the (log) likelihood ratio statistic for testing independence of rows and columns may be written in the form

$$G^2 = -2\log\frac{L_{H_0}}{L_{H_1}} = 2\sum o\log\left(\frac{o}{e}\right)$$

   where o and e denote observed frequency and fitted frequency respectively.

   What is the asymptotic distribution of G^2 under the null hypothesis that rows and columns are independent?

6. The following table gives the numbers of children who were nasal carriers or non-carriers of *Streptococcus pyogenes*, classified by size of tonsils.

	Size of tonsils		
	Not enlarged (but present)	Enlarged	Greatly enlarged
Carriers	19	29	24
Non-carriers	497	560	269

   Assuming a multinomial distribution for the frequencies, use the G^2 statistic to test the null hypothesis of independence of presence/absence of *Streptococcus pyogenes* and size of tonsils.

7. Show that the likelihood ratio statistic $G^2 = 2\sum o\log\frac{o}{e}$ is approximately equal to the χ^2 goodness-of-fit statistic $X^2 = \sum\frac{(o-e)^2}{e}$. [Hint: use log series expansion.]

   Compare the values of X^2 and G^2 for the contingency table given in Question 6.

8. Let $Y_1, \ldots, Y_n$ be independent, Y_i having a Poisson distribution with mean $\lambda_i = \theta_1 + \theta_2 x_i$ for fixed x_i. The unknown parameter vector $\theta = (\theta_1, \theta_2)^T$ is to be estimated from data.

   Show that the log likelihood function, up to an additive constant, is given by

$$l(\theta_1, \theta_2) = -\sum_{i=1}^{n}(\theta_1 + \theta_2 x_i) + \sum_{i=1}^{n} Y_i \log(\theta_1 + \theta_2 x_i)$$

   Obtain Fisher's method of scoring for this problem. Give an approximate expression for the variance covariance matrix of $\widehat{\theta}$.

9. The following discrete data on textile faults, which has $n = 32$ measurements of cloth length (x) and the number of faults (y), can be described by the Poisson regression model given in Question 8.

x:	551	651	832	375	715	868	271	630	491	372	645	441	895	458	642	492
y:	6	4	17	9	14	8	5	7	7	7	6	8	28	4	10	4
x:	543	842	905	542	522	122	657	170	738	371	735	749	495	716	952	417
y:	8	9	23	9	6	1	9	4	9	14	17	10	7	3	9	2

(a) Complete one iteration of the method of scoring by hand. Take the starting value as $(\theta_1, \theta_2)_0 = (0, 1)$.

(b) Write MAPLE/R/Python/MATLAB/etc code to implement the method of scoring for this problem, then use it to determine the maximum likelihood estimates of θ_1 and θ_2.

(c) Provide asymptotic standard errors for the corresponding estimators.

10. For the exponential family density

$$f(y;\theta) = \exp[y\theta - c(\theta) + d(y)]$$

define the function $g(\cdot)$ to be the inverse function of $c'(\cdot)$, i.e. $\theta = g(\mu)$ iff $\mu = c'(\theta)$. Let $z = g(y)$. Show that if $\text{var}(Y)$ is small, then $\text{var}(Z) \simeq \frac{1}{\text{var}(Y)}$.

11. Suppose Y has an exponential distribution with mean μ. Does the density of Y belong to the exponential family defined in Question 10? If so, what is the canonical parameter θ as a function of μ?

For a single observation consider the generalized linear model (with non-canonical link)

$$\eta = \log(\mu) = \mathbf{X}'\beta.$$

Show that the score is

$$\frac{\partial l}{\partial \beta_j} = (y-\mu)\frac{\partial \theta}{\partial \eta}x_j = (y-\mu)\frac{1}{\mu}x_j$$

and that the information matrix has elements

$$-E\left[\frac{\partial^2 l}{\partial \beta_k \partial \beta_j}\right] = x_k x_j \text{var}(Y)\left(\frac{\partial \theta}{\partial \eta}\right)^2$$

Explain why for this model the iterative weighted least squares algorithm would involve a series of **unweighted** least squares regressions.

12. Use the R function `lmer` to fit a model to the data below from Example 'Variation in the yield of a dyestuff' in the document 'Generalized Linear Mixed Models (GLMM)'.

	Batch					
	1	2	3	4	5	6
Yields	1545	1540	1595	1445	1595	1520
	1440	1555	1550	1440	1630	1455
	1440	1490	1605	1595	1515	1450
	1520	1560	1510	1465	1635	1480
	1580	1495	1560	1545	1625	1445

13. Use the R function `glmer` — see code below — to fit a model for Example 'Contagious bovine pleuropneumonia (CBPP)' in the document 'Generalized Linear Mixed Models (GLMM)'.

```
library(lme4)
## response as a matrix
(m1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
family = binomial, data = cbpp))
## response as a vector of probabilities and usage of argument "weights"
m1p <- glmer(incidence / size ~ period + (1 | herd), weights = size,
family = binomial, data = cbpp)
## Confirm that these are equivalent:
stopifnot(all.equal(fixef(m1), fixef(m1p), tolerance = 1e-5),
all.equal(ranef(m1), ranef(m1p), tolerance = 1e-5))checkConv
13
## GLMM with individual-level variability (accounting for overdispersion)
cbpp$obs <- 1:nrow(cbpp)
(m2 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd) + (1|obs),
family = binomial, data = cbpp))
```