MATH3823 weekly exercises

Dr Stuart Barber

March 2022

These questions are intended to be worked on alongside reading the notes. To help keep track of where we are up to, they are separated by week they were set.

Week 2

1. Verify that if $q = 1/(1 + e^{-x})$ then $x = \log(q/(1 - q))$.
2. For the birthweight data example in section 1.2:
   - What obvious model have we not listed in the notes? Looking at the data plots in Figure 1.2, comment on whether you think this model is likely to be worth considering
   - Test whether the `age * sex` interaction term is statistically significant for these data and comment on the resulting model.

Week 3

3. Express each of the following distributions as an exponential family distribution, possibly with a scale parameter. Use the properties of exponential families to find the expectation and variance of each distribution
   - The geometric distribution with probability mass function

     $$f(y) = (1 - p)^{y-1}p; \qquad y = 1, 2, 3 \ldots$$

     where $0 < p < 1$.
   - The gamma distribution with probability density function

     $$f(y) = \Gamma(\alpha)^{-1}\lambda^{\alpha}y^{\alpha-1}e^{-\lambda y} \qquad y \geqslant 0,$$

     where $\alpha, \lambda > 0$. (Hint: Treat $1/\alpha$ as a scale parameter and let θ be a suitable function of λ and α.)

Week 4

4. Find the canonical link functions for the geometric and gamma distributions.

Week 5

5. Check that you can derive the formulae given for deviance of normal, binomial, and Poisson models just after equation (2.63) in the lecture notes.
6. (After reading section 2.6 of the notes) Go back to the birthweight example and consider how the deviance tests from section 2.6 relate to the tests we used to choose which model was the best fit to the data.

Week 6

7. Complete the exercises in the R script from lecture 4:
   - Check the two different ways of fitting a Poisson GLM to the toy data give the same result;
   - Check whether a smaller model is adequate for the carbohydrate uptake in diabetes data set;
   - Fit alternative link functions to the beetle data and plot the results.
8. Use the fitted logisitic regression model to predict what dose of gaseous carbon disulphide would kill 90% of beetles.

Week 8

9. Analyse the colon cancer data available in Minerva. Specifically, use a logistic regression model to investigate whether the genetic mutation and/or the environmental exposure affect the incidence of colon cancer.

   R hints:

   - encode 'agecat' as a factor, as shown in the 'data' folder on Minreva.
   - Note that the response variable for a binomial glm in R can be a vector (not a matrix) in the special case of $m = 1$, so a formula of the form 'case $\sim \ldots$.' will work for this data set.

Week 11

10. Overdispersion is an occasional problem when fitting generalised linear models with a known scale parameter in situations where there is unexplained variation. In this

exercise we illustrate the problem in Poisson regression. The starting point is the observation that if $Y \sim P(\lambda)$, then $E(Y) = \lambda$ and $\text{var}(Y) = \lambda$.

a. Consider joint random variables (X, Y) where X takes two possible values with equal probabilities, $\Pr(X = 1) = \Pr(X = 2) = 1/2$. Suppose the conditional distribution of Y given X is Poisson,

$$Y|X = 1 \sim P(\lambda_1), \quad Y|X = 2 \sim P(\lambda_2), \qquad (*)$$

where $\lambda_1 < \lambda_2$. Let $\lambda = (\lambda_1 + \lambda_2)/2$ denote the average value. Thus the marginal distribution of Y is a mixture of two Poisson distributions. Show that

$$E(Y) = \lambda, \quad \text{var}(Y) = \lambda + (\lambda_1 - \lambda_2)^2/4,$$

that is, although the mean of Y is the same under the mixture (or conditional Poisson) model as under the Poisson model, the variance is larger.

b. This phenomenon might be observed in data as follows. Let $n = 60$ and let an explanatory variable x_i take the value $x_i = 1$ for $i = 1, \ldots, 30$ and $x_i = 2$ for $i = 31, \ldots, 60$. Suppose that the observations $y_i|x_i$ come from the above conditional Poisson model $(*)$.

Consider fitting the following two models in R with Poisson errors and a log link function:

$$\text{(i)} \;\; y \sim 1, \qquad \text{(ii)} \;\; y \sim x.$$

Since model (ii) is the correct model, it should yield a good fit to the data. But if the experimenter does not know about the variable x, it will only be feasible to fit model (i). Let $\overline{Y}$ and S^2 denote the sample mean and variance of the $\{Y_i\}$, $i = 1, \ldots, 60$. Show that

$$E(\overline{Y}) = \lambda, \quad E(S^2) = \lambda + \frac{60}{59}(\lambda_1 - \lambda_2)^2/4.$$

Hence show that the χ^2 goodness of fit statistic for model (i) will indicate a poorly fitting model if λ_1 and λ_2 are far apart.

c. This example is very simple, but overdispersion can occur much more widely. Why is overdisperson not a problem for genereralised linear models in which the response distribution includes a scale parameter?

11. (Based on Dobson & Barnett, pp 144–145). Suppose there are 2 groups of people: the first group is exposed to some pollutant and the second group is not. In a prospective study, each group is followed for several years and categorized according to the presence or absence of some disease. Let π_i denote the probability that a person in group i contracts the disease, $i = 1, 2$. The following 2×2 table summarizes the different possibilites.

	Diseased	Not diseased
Exposed	π_1	$1-\pi_1$
Not exposed	π_2	$1-\pi_2$

Note that the sum of each row is 1. For each $i = 1, 2$, the *odds* of contracting the disease is defined by

$$O_i = \pi_i/(1-\pi_i),$$

and a comparision between these two probabilities is given by the *odds ratio*

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}.$$

a. Show that $\phi = 1$ if and only if there is no difference between the control and exposed groups. What does it mean if $\phi > 1$?

b. Consider now m 2×2 tables of this form, $j = 1, \ldots, m$, with probabilities π_{ij} represented by a logistic model

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta_i x_j, \ i = 1, 2, \ j = 1, \ldots, m,$$

where x_j is some specified quantitative explanatory variable. Interpret the parameters α_i and β_j, and give their effect on the log odds ratio $\log \phi_j$, say, for each table. Show that $\log \phi_j$ is constant across the m tables if $\beta_1 = \beta_2$.

c. Give a practical example where such a model might be appropriate.

d. How would you express this model in the R computer language?

12. Consider a $2 \times m$ contingency table with entries y_{ij}, $i = 1, 2$, $j = 1, \ldots, m$. The rows label STATUS (= 1,2 for alive/dead) and the columns label AGE groups (= 1,. . . ,m). Consider a Poisson model $P(\lambda_{ij})$ with

$$\log(\lambda_{ij}) = \delta + \alpha_i + \beta_j + \gamma_i j$$

and discuss suitable aliasing conditions on the parameters (note that j is treated as a quantitative variable in the last term). Show that this model, conditioned on the column totals y_{+j}, reduces to a product binomial model $B(y_{+j}, \pi_j)$, $j = 1, \ldots, m$, and find the form of π_j. Which parameters are estimable under the binomial model.

13. (From Dobson & Barnett, p 163) This question should be done in a computer package such as R. You should think carefully about which variables, if any, to condition on in your analysis: HOME = 1,2,3, CONTACT = 1,2, or SATISFACTION = 1,2,3.

The data relate to an investigation into satisfaction with housing conditions in Copenhagen. Residents of selected areas living in rented houses built between 1960 and 1968 were questioned about their satisfaction and their degree of contact with other residents. The data were tabulated by type of housing. Investigate the associations between satisfaction, contact with other residents and type of housing.

Low Contact:

Satisfaction:	Low	Medium	High
Tower blocks	65	54	100
Apartments	130	76	111
Houses	67	48	62

High Contact:

Satisfaction:	Low	Medium	High
Tower blocks	34	47	100
Apartments	141	116	191
Houses	130	105	104

a. Produce appropriate tables of percentages to gain initial insights into the data; for example, percentages in each contact level by type of housing and level of satisfaction, or percentages in each level of statisfaction by contact and type of housing.
b. Using e.g. R, fit various log-linear models to investigate interactions between the variables.
c. For some model that fits (at least moderately) well, calculate the Pearson residuals and use them to find where the largest discrepencies are between the observed and expected values.