

- (1) Brown, B.W. (1980) describes a study of 53 prostate cancer patients. Five binary predictor variables ('aged', 'stage', 'grade', 'xray' and 'acid') were measured before surgery. For these variables 0 indicates the risk factor is absent, and 1 indicates that it is present. The patients then had surgery to determine whether there was nodal involvement (NI equal 1) or not (NI equal zero) in the cancer. The aim was to find out which predictor variables were most important. The table below shows the first and last 3 rows of the data.

	NI	aged	stage	grade	xray	acid
1	1	0	1	1	1	1
2	1	0	1	1	1	1
3	1	0	1	1	1	1
...						
51	1	0	0	1	0	1
52	0	0	0	0	1	1
53	0	0	0	0	1	0

- (a) Specify a generalised linear model for these data using  $NI$  as the binary response, and 'aged', 'stage', 'grade', 'xray' and 'acid' as binary explanatory variables, using the canonical link function for your model.
- (b) Define the saturated model for the GLM you gave in the previous part, and calculate the residual deviance (as a function of parameter MLEs).
- (iii) Fitting a GLM with  $NI \sim 1 + xray + acid + stage + aged + grade$  gave the following residual deviances:

	Resid. Dev
Intercept	70.25
xray	60.93
acid	54.79
stage	49.18
aged	48.76
grade	47.61

Resid.Dev in row  $i$  is the residual deviance for a GLM model with a linear predictor including all the variables in rows 1 to  $i$ . For example, the GLM with linear predictor  $NI \sim 1 + xray + acid$  has RD equal 54.79. Give the null deviance and test for any linear dependence. Calculate the AIC for each of these nested models. Carry out model selection using chisq tests from the change in residual deviance and again *via* the AIC. Explain why we should not test for goodness of fit by directly comparing the residual deviance (of the final model) to a  $\chi^2$  variate.

- (2) (*Poisson GLM, data from Dr. Paul D. Baxter lecture notes 'Generalised Linear Models by Example'*) The data for this question are in the file `aids.csv` (a spreadsheet file) at the BS1a website. The columns of this data set are
- **cases**, the number of AIDS cases in a given 3 month period (a 'quarter') from Jan 1983 to Mar 1994 as reported by the Public Health Laboratory Service, Communicable Disease Surveillance Unit, London.
  - **qrt**, the quarter of the year
  - **date** the year in fractions of a quarter

The aim of the analysis is to see if there are any seasonal effects, to see if an exponential increase in the number of aids cases over the period is a reasonable model, and measure the exponential growth rate.

```
> aids<-read.csv('aids.csv')
> attach(aids)
> names(aids)
[1] "X"      "cases"  "qrt"    "date"
> qrt<-as.factor(qrt)
> plot(date,cases)
```

The levels of `qrt` are 1, 2, 3, 4. Consider fitting a model with  $y_i = \text{cases}[i]$ ,  $x_{i,2}$ ,  $x_{i,3}$  and  $x_{i,4}$  dummy indicator variables for levels 2, 3 and 4 of `qrt[i]`, and  $x_{i,5} = \text{date}[i]$ . The linear predictor is

$$\eta_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5},$$

and the stochastic part of the GLM is  $Y_i \sim \text{Poisson}(\mu_i)$ .

- (a) Give a brief mathematical explanation of what it means to use a log link function, or a square-root link function.
  - (b) Using R, and the `family = poisson` option to `glm()`, fit the model. Try the log and square-root link functions (use `glm(cases~date, family=poisson(link=sqrt))` to get the square-root link function).
  - (c) Is `qrt` a significant explanatory variable? Briefly interpret the estimated parameters of the regression.
  - (d) Make a test for goodness of fit and comment on your findings.
- (3) Consider the data in Table 1.

	recovered		
	Y	N	
treated	$r_t$	$d_t$	$n_t$
untreated	$r_u$	$d_u$	$n_u$
	$r$	$d$	$n$

TABLE 1. Counts for treated (cases) and untreated (controls) individuals recording positive response to treatment.

- (a) (*prospective*) Take a group of  $n$  patients with the same sickness. Treat  $n_t$  (the cases) and leave  $n_u$  untreated (the controls). Record the numbers (see the table) of patients who recover in each class. Let

$$\pi_t = \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}$$

be the probability for a treated patient to recover and let

$$\pi_u = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

be the probability for an untreated patient to recover. Calculate the MLE for the log-odds ratio for recovery for treated and untreated cases.

- (b) (*retrospective*) Suppose we look through hospital records and find  $R$  patients who recovered and  $D$  patients who didnt. Select at random  $r$  from the  $R$  and  $d$  from the  $D$ . From these selected patients, record the numbers receiving treatment. Let

$\Pr(\text{recovered}|\text{treated, selected})$  be the probability that a patient recovers, given that they were selected for the retrospective study, and had been treated. Show that in this second study

$$\Pr(\text{recovered}|\text{treated, selected}) = \frac{\exp(\alpha' + \beta)}{1 + \exp(\alpha' + \beta)},$$

(same  $\beta$  as before) and calculate  $\alpha'$  in terms of  $\alpha$ .

(c) Hence, show that the MLE's for the log-odds ratio for recovery under the two treatments is the same in prospective and retrospective studies.

(4) (*extra for experts*) Consider logistic regression with  $n$  binary observations  $y_1, \dots, y_n$  and linear predictors  $\eta_i = \mathbf{x}_i^T \beta$   $i = 1, 2, \dots, n$ . Show that the likelihood can be written

$$L(\beta; y) = \prod_i \left( \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{y_i} \left( \frac{1}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{1-y_i}.$$

Consider the data  $y = (0, 0, 0, 0, 1, 1, 1, 1)$ ,  $x = (0, 1, 2, 3, 4, 5, 6, 7, 8)$  (and design matrix  $X = (1, x)$  with  $p = 2$  variables). Explain why the MLE is not defined. Can you generalise this result to  $p > 2$ ?