STATISTICAL MODELLING Part IIC.

Example Sheet 4 (of 4) RN/Lent 2011

1. (Weighted least squares) Let $Y_1, \ldots, Y_n$ be independent, with $Y_i \sim N(\mu_i, \sigma_i^2)$, where $\mu_i = x_i^T \beta$ and $\sigma_i^2 = \sigma^2 a_i$, with σ^2 unknown, but $a_1, \ldots, a_n$ known. Show that the maximum likelihood estimator $\hat{\beta}$ is the solution to the weighted least squares problem of minimising $(Y - X\beta)^T W (Y - X\beta)$, where W and X should be specified.

   Deduce that $\hat{\beta}$, which is also called the weighted least squares estimator, satisfies $\hat{\beta} = (X^T W X)^{-1} X^T W Y$.

2. (Iterated weighted least squares) Recall that the mth iteration of the Fisher scoring algorithm in a generalised linear model is

$$\hat{\beta}_m = \hat{\beta}_{m-1} + i(\hat{\beta}_{m-1})^{-1} U(\hat{\beta}_{m-1}).$$

   Let $\hat{Z}_{m-1} = (\hat{Z}_{m-1,1}, \ldots, \hat{Z}_{m-1,n})^T$, where $\hat{Z}_{m-1,i} = \hat{\eta}_{m-1,i} + (Y_i - \hat{\mu}_{m-1,i}) g'(\hat{\mu}_{m-1,i})$, with $\hat{\eta}_{m-1,i} = (X\hat{\beta}_{m-1})_i$ and $\hat{\mu}_{m-1,i} = g^{-1}(\hat{\eta}_{m-1,i})$, for $i = 1, \ldots, n$. From the expressions for $U(\beta)$ and $i(\beta)$ computed in Ex. Sheet 3, question 8, deduce that

$$\hat{\beta}_m = (X^T \hat{W}_{m-1} X)^{-1} X^T \hat{W}_{m-1} \hat{Z}_{m-1},$$

   where $\hat{W}_{m-1}$ is a matrix which you should specify.

3. Consider a generalised linear model with Poisson responses and the canonical link function, with linear predictor $\eta = (\eta_1, \ldots, \eta_n)^T$ given by $\eta_i = \alpha + x_i^T \beta$, for $i = 1, \ldots, n$. Argue that the deviance may be approximated by Pearson's χ^2 statistic. *Hint: if stuck, Taylor expand.*

4. (Short Tripos 2005/2/5I) Below are three `R` commands, and the corresponding output (which is slightly abbreviated). Explain the effects of the commands. How is the deviance defined, and why do we have d.f.=7 in this case? Interpret the numerical values found in the output.

```
> n <- c(3,5,16,12,11,34,37,51,56)
> i <- c(1,2,3,4,5,6,7,8,9)
> summary(glm(n~i,poisson))
  deviance = 13.218
      d.f. = 7
  Coefficients:
                 Value     Std.Error
  (intercept)    1.363     0.2210
  i              0.3106    0.0382
```

5. Let $Y = (Y_1, \ldots, Y_m)$ be a random vector having independent components, with $Y_i \sim \text{Poi}(\mu_i)$ for $i = 1, \ldots, m$. Show that, conditional on $\sum Y_i = n$, we have that $Y \sim \text{Multi}(n; p_1, \ldots, p_m)$, where $p_i = \mu_i / \sum \mu_j$ for $i = 1, \ldots, m$.

6. (a) In a two-way contingency table, consider the hypothesis that the row index and column index of an observation are independent. Write down the mulitnomial and surrogate Poisson models corresponding to this hypothesis. Show that the fitted values from the surrogate model (so also from the multinomial model) have the same row and column totals as the observed values.

   (b) Now consider a two-way contingency table in which the row totals are fixed and we are interested in a hypothesis of the homogeneity of the different rows. Write down the multinomial and surrogate Poisson model corresponding to this hypothesis and argue again that the fitted row and column totals respect the observed values.

   In each case, explain how would you test these hypotheses, assuming that the fitted values in each cell are not too small.

7. The data below come from a study of hypertension (high blood pressure), obesity and alcohol intake in Western Australia. The alcohol categories are in 'drinks' per day. Read the data into `R` and define appropriate factors using `gl`. Think about questions of interest for this data and fit appropriate models to study these questions. What are your conclusions?

		Alcohol Intake			
Obesity	BP	0	1-2	3-5	6+
Low	Yes	5	9	8	10
Low	No	40	36	33	24
Average	Yes	6	9	11	14
Average	No	33	23	35	30
High	Yes	9	12	19	19
High	No	24	25	28	29

8. (Long Tripos 2005/4/13I)

   (a) Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that Y_1 has probability density function

$$f(y_i|\beta, \nu) = \left(\frac{\nu y_i}{\mu_i}\right)^{\nu} e^{-y_i \nu/\mu_i} \frac{1}{\Gamma(\nu)} \frac{1}{y_i} \quad \text{for } y_i > 0$$

   where

$$1/\mu_i = \beta^T x_i \,, \text{ for } 1 \leq i \leq n,$$

and $x_1, \ldots, x_n$ are given p-dimensional vectors, and ν is known.
Show that $\mathbb{E}(Y_i) = \mu_i$ and that $\mathrm{var}(Y_i) = \mu_i^2/\nu$.

(b) Find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, and suggest an iterative scheme for its solution.

(c) If $p = 2$, and $x_i = \begin{pmatrix} 1 \\ z_i \end{pmatrix}$, find the large-sample distribution of $\hat{\beta}_2$. Write your answer in terms of a, b, c and ν, where a, b, c are defined by

$$a = \sum \mu_i^2, \quad b = \sum z_i \mu_i^2, \quad c = \sum z_i^2 \mu_i^2.$$