STATISTICAL MODELLING Part IIC.

Example Sheet 4 (of 4) BS/Lent 2013

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. 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

7. (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 $\text{var}(Y_i) = \mu_i^2/\nu$.

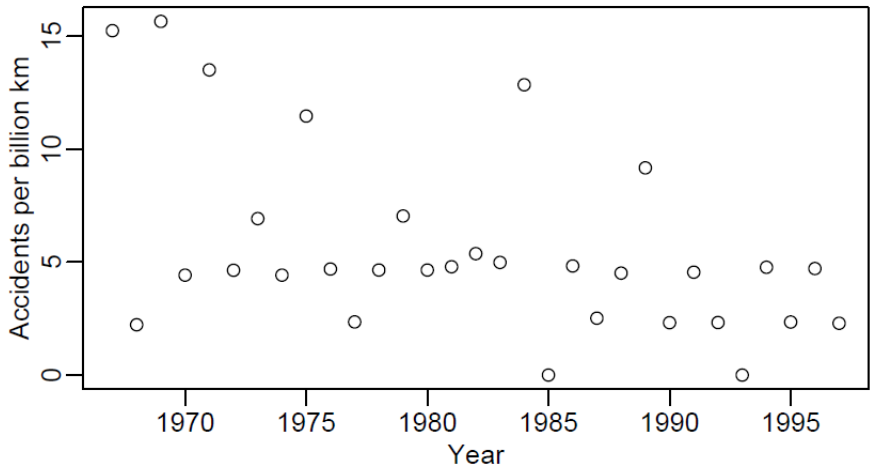
   (b) Find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, and suggest an iterative scheme for its solution.

   (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

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

8. The number of fatal train accidents were recorded during the years 1967 to 1997 in the United Kingdom. Some years there were no fatal accidents at all while

in 1967 and again in 1969, there were seven such accidents. The total number of kilometres (in billions) traveled each year by all trains was recorded. The accidents per billion kilometers is plotted in the following Figure. A poisson GLM was fit with the model formula:




```
Accidents ~ offset(log(Train.km)) + Year
```

where `Accidents` is the number of accidents, `Train.km` is the billion kilometres traveled by trains and `Years` runs from `1` to `31` representing the years 1967 to 1997. The following output was obtained:

```
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   2.3211       0.2039    11.4    <2e-16
Year         -0.0414       0.0134    -3.1    0.0019

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 37.645 on 30 degrees of freedom
Residual deviance: 27.599 on 29 degrees of freedom
```

(a) About how many billions of kilometres were traveled by UK trains in 1967?

(b) The statistic -3.1 can be used to test for a linear trend in year. Compute a different statistics that can be used to test the same hypothesis.

(c) Suppose we plotted the fit of the model on the data as depicted in Figure. Would the fit appear as a straight line or a curve? Explain.

(d) Train accidents can be divided into three types - those caused by the driver ignoring a warning signal, those caused by some other driver error and those that are not due to driver error. A multinomial logit model with these three response categories and year as a predictor was fit to the data and the following output obtained:

```
Coefficients:
                (Intercept)          Year
Other.Driver      -0.51821      -0.064898
Non.Driver         0.26333       0.014835

Residual Deviance: 137.76
```

In 1967, this model predicts one of the three types of accident to be the most common. Does this accident type remain the most common predicted type over the years of the study? Explain.

(e) The null deviance was 140.42. Construct an appropriate statistic for testing a time trend. Give the degrees of freedom for this test statistic.

(f) Suppose you wanted to predict the number of non-driver preventable fatal accidents in 1998. What one additional numerical piece of information, other than those presented above, would you need to complete the computation?

9. At the end of 1991, the US Fish and Wildlife service conducted a survey of 196 bass anglers in North and South Carolina to determine their sensitivity to the coast of such fishing trips. Subjects were asked whether they would have fished that year if the cost of the trips had been increased by an amount of dollars specified by the interviewer. The variables were:

`yes` yes = 1, no = 0 - answer to "Would you still have fished?"
`cost` proposed additional cost of trips
`catch` number of bass caught during year
`employed` employed or not
`education` in years
`married` married or not
`sex` Female or Male
`age` in years
`nc` North Carolina = 1, South Carolina = 0

A binomial GLM was fit with `yes` as the response. The following summary output was obtained:

```
Deviance Residuals:
   Min       1Q    Median      3Q       Max
-2.192   -1.026     0.496   0.962     2.075

Coefficients:
                       Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)            2.682906     1.125233     2.38    0.0171
cost                  -0.001752     0.000562    -3.12    0.0018
catch                  0.005895     0.002249     2.62    0.0088
income                 0.014585     0.009916     1.47    0.1413
employed-NotEmployed  -1.432589     0.603035    -2.38    0.0175
education             -0.082740     0.062654    -1.32    0.1866
Married-NotMarried    -0.350022     0.405820    -0.86    0.3884
SexMale               -1.003908     0.477927    -2.10    0.0357
age                   -0.005469     0.015808    -0.35    0.7294
nc                    -0.441621     0.325095    -1.36    0.1743

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 270.06 on 195 degrees of freedom
Residual deviance: 228.44 on 186 degrees of freedom
AIC: 248.4

Number of Fisher Scoring iterations: 3
```

(a) Give an interpretation of raising the cost of fishing trips by $100.

(b) There is a difference of 41.56 in the deviances in the output above. What hypothesis does this statistic test and what conclusion should then be made?

(c) Using this model, predict how much additional cost would need to be imposed to make it certain no subject in the study would want to go on a fishing trip.

(d) If all other predictors were held constant, what does the model say about the fishing preferences of men compared to women?

(e) Are there any outliers? Explain.

(f) What would be the best way to determine the statistical significance of the North vs. South Carolina effect in this model?