STATISTICAL MODELLING

Example Sheet 3

(Warning: somehow most people find questions 2, 3 to be hard. There is a practical point behind these 2 questions – and the others too!)

1. The observed waiting times $t_1, \ldots, t_n$ are independent, with T_i having pdf

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

where ν is a known parameter, and μ_i depends linearly on a known covariate x_i through the following link function:

$$\frac{1}{\mu_i} = \alpha + \beta x_i, \quad \text{for } t_i > 0.$$

Let $a_1 = \Sigma t_i, a_2 = \Sigma x_i t_i$. Show that (a_1, a_2) are sufficient for (α, β).

The observations $t_1, \ldots, t_m$ are times between consecutive earthquakes in Mexico City, and the observations $t_{m+1}, \ldots, t_n$ are the times between consecutive earthquakes in Turkey. Take $x_1 = \ldots = x_m = 0$ and $x_{m+1} = \ldots = x_n = 1$, and discuss the estimation of β when m, n are large, quoting any general asymptotic likelihood results needed for your solution.

This introduces you to another 'error distribution' available in glm: the gamma.

2. Your client gives you data

$$(y_{ij}, x_{ij}, 1 \le j \le n_i, 1 \le i \le t)$$

and asks you to fit the model

$$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij},$$

with $\epsilon_{ij} \sim NID(0, \sigma^2)$, σ^2 unknown. He has arranged the experiment so that

$$x_{ij} = x_i, \quad 1 \le j \le n_i, 1 \le i \le t.$$

Find expressions for $(\hat{\alpha}, \hat{\beta})$, the least squares estimators.

Your client now observes that a consequence of the model above is that

$$\mathbb{E}(\overline{y}_i) = \alpha + \beta x_i, \quad 1 \le i \le t, \quad (\text{where } \overline{y}_i = \sum_j y_{ij}/n_i).$$

He suggests that some of your (highly paid) time could be saved by reading in the data as the t pairs $(\overline{y}_i, x_i)$, $1 \le i \le t$ instead of the original $(n_1 + \cdots + n_t)$ pairs of points. How do you advise him? Give reasons for your answer.

[Hint: write down the likelihood given by
a) the full set of data (y_{ij}, x_{ij}), and
b) the reduced set of data $(\overline{y}_i, x_i)$.
Show that the maximum likelihood estimates of (α, β) are the same for a) and for b).]

3. Another client gives you data

$$(0, x_1), (1, x_1), (0, x_1), (1, x_2), (1, x_2),$$
$$(1, x_2), (0, x_2), (1, x_3), (1, x_3), (0, x_3),$$
$$(1, x_4), (1, x_4), (1, x_4), (1, x_4), (0, x_4),$$

i.e. 15 independent 'units', with the first digit of each pair being 1 or 0 with probabilities $p(x), q(x)$ respectively. She asks you to fit the model

$$\log(p(x)/q(x)) = \alpha + \beta x$$

and to use the appropriate difference in deviances to test the hypothesis $\beta = 0$.

You observe that the data can in fact be compressed and read in as four independent values $(1, 3, x_1), (3, 4, x_2), (2, 3, x_3), (4, 5, x_4)$.
Would this approach result in misleading your client? Give reasons for your answer.
[Hint: think LIKELIHOODS, as in question 2 above.]

4. Suppose

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

where Y is $n \times 1$, X is $n \times p$ and of rank p, and β is the unknown vector of parameters.

(a) Show that $\hat{\beta} = (X^T X)^{-1} X^T Y$ is the lse of β.

(b) Show that $\hat{Y} = X\hat{\beta} = HY$ say, where $H = H^T$ and $HH = H$ ($\hat{Y}$ is the vector of 'fitted' values).

(c) Suppose $e = Y - \hat{Y}$ (the vector of *residuals*). Show that

$$e \sim N(0, \sigma^2(I - H))$$

and hence

$$e_i \sim N(0, \sigma^2(1 - h_i)), \quad 1 \le i \le n$$

where h_i is the i^{th} diagonal element of H.

(d) Show that if λ is an eigenvalue of H, then λ is either 1 or 0. Hence show $\sum_1^n h_i = p$. (The quantity h_i is called the 'influence' or 'leverage' of the i^{th} data point x_i where $y_i = \beta^T x_i + \epsilon_i$, $1 \le i \le n$.)

(e) Show that $0 \le h_i \le 1$, and find h_i for simple linear regression, i.e. for

$$y_i = \beta_1 + \beta_2(x_i - \overline{x}) + \epsilon_i.$$

(f) How do you interpret the statement "Values of x_i corresponding to large leverage exert a pronounced effect on the fit of the linear model at (x_i, y_i)"?

(g) Use R to construct leverages (also called 'influence values') and qplots for simple (y, x) data-sets.

5. (a) Suppose $(y_i) \sim Mn(n, (p_i))$, $\sum_1^k p_i = 1$. Consider testing

$$H_0 : \log p_i = \mu + \beta x_i, \quad 1 \le i \le k$$

where $\mathbf{x}$ is given, and μ is s.t. $\Sigma p_i = 1$, thus

$$e^\mu = 1/\Sigma e^{\beta x_i}.$$

Show that $\hat{\beta}$ is the solution of

$$(*) \qquad \Sigma y_i x_i = n \Sigma x_j e^{\beta x_j} / \Sigma e^{\beta x_j}.$$

Let $e_i = np_i(\hat{\beta})$ ('expected values under the null hypothesis').

(b) Now suppose instead that

$$y_i \sim \text{independent Poisson } (\mu_i), \quad \mu_i \ge 0.$$

Consider testing

$$HP_0 : \log \mu_i = \mu' + \beta x_i, \quad 1 \le i \le k.$$

Find $\ell(\beta, \mu')$ the loglikelihood function, and hence show that the mle for β is given by $\hat{\beta}$, as in equation $*$ above. Show also that

$$\sum_1^k \mu_i(\hat{\beta}, \hat{\mu}') = y_+$$

(i.e. observed and expected values of y_+ agree exactly at mle). Comment on the glm application of the Poisson distribution for problem (a).

6. A random sample of 193 individuals, classified according to four 2-level factors, gave the following $2 \times 2 \times 2 \times 2$ contingency table:

		$B=1$		$B=2$	
		$A=1$	$A=2$	$A=1$	$A=2$
$S=1$	$D=1$	89	2	4	1
	$D=2$	8	4	3	1
$S=2$	$D=1$	70	6	2	0
	$D=2$	1	0	1	1

Let $p_{ijk\ell}$ denote the corresponding underlying cell probabilities where i, j, k, ℓ correspond to the factors S, D, B and A respectively. With the Poisson distribution for n, the cell frequencies, and log link, glm (in S-Plus) finds, in the following order,

(a) $n \sim S * D * A * B$ gives deviance $= 0$, df $= 0$,

(b) $n \sim (S + D + A + B) \wedge 3$ gives deviance 2.72, df $= 1$,

(c) $n \sim (S + D + A + B) \wedge 3 - S : A : B - D : A : B$ gives deviance 3.42, df $= 3$,

(d) $n \sim (S + A + B) * D$ gives deviance 8.48, df $= 8$.

Write down the model for $p_{ijk\ell}$ for each of (a), (b), (c) and (d), give an interpretation of the deviance at each stage, and give an interpretation in terms of conditional independence for the model in (d).

How would you check the fit of the model S, D, A, B mutually independent?

7. *Which?* (August 1980, p. 436) gives the data in the table below on lager (available to you on catam stats). The columns are price per half pint, o.g. (original gravity), percent alcohol, calories per half pint, and 'experts' rating'. The original gravity is described by *Which?* as "another guide to strength; it's a measure of what has gone into the beer besides water, and is used to calculate the duty payable". The 'experts' rating' column contains entries 3, 4, 5 for lagers actually tasted (5 being the most liked and 3 the least liked), and an entry 0 for lagers not actually tasted.

Reading the data provided on file, use lm() to answer the following questions:

(a) Does the price depend on o.g., percent alcohol, and calories per half pint, and if so, how?

(b) Is the price of those lagers tasted significantly different from the price of those not tasted?

Price	o.g.	percent	cal	experts' rating
17.5	1031	3.1	76	3
20.5	1032	3.2	79	3
22.5	1031	3.3	76	3
22.5	1032	3.3	78	0
18.5	1035	3.4	83	4
22.5	1033	3.5	78	3
23	1031	3.6	76	3
22	1033	3.6	82	3
19.5	1033	3.6	81	0
18.5	1033	3.6	81	0
22.5	1036	3.7	88	0
22	1034	3.7	83	3
24	1036	3.8	87	5
24	1036	3.8	87	0
18.5	1037	3.8	91	4
19.5	1038	3.9	91	5
25	1041	4.0	100	3
26	1036	4.0	84	3
20	1037	4.0	90	5
27	1037	4.0	91	5
22.5	1037	4.1	89	3
20.5	1038	4.1	92	0
43.5	1045	4.7	110	4
27.5	1045	4.8	109	4
29	1046	4.8	110	4
26.5	1047	4.9	116	3
29	1046	4.9	112	0
24.5	1047	4.9	116	0
18.5	1034	3.5	81	3
31	1045	5.0	109	0
32	1047	5.0	117	3
24	1046	5.0	111	4
29	1046	5.1	111	3
33	1046	5.1	110	0
29	1048	5.2	119	0
33.5	1050	5.4	121	0
26	1051	5.5	125	5
43	1058	6.0	146	0
31.5	1079	8.9	197	0
31.5	1081	8.9	204	0

8. In a study on the possible relationship between the psychological well-being of mothers and that of their children, the psychiatrist Prof I. Goodyer collected data, some of which is summarised below. Table 1 shows, for the 200 children in the study, how many in each of eight categories were 'cases', i.e. anxious or depressed, and the total number in each category. Those who are not cases are 'controls', assumed well. The eight categories are defined by three binary factors:

(a) 'rmq', which corresponds to a particular measure of the mother's psychological well-being;

(b) 'mcr', which indicates whether or not the mother has good 'confiding' relations with other adults;

(c) 'events', which indicates whether or not the child has experienced recent stressful life events in the 12 months prior to the study.

In each case a value of 1 for the factor corresponds to its status being 'good' or 'normal', and a value of 2 corresponds to its being 'poor'.

Assuming that the number of cases in a given category is binomial, given the total in that category, find a model relating the probability that a child is a case rather than a control to the three factors given. Discuss carefully how to interpret your best-fitting model, and its estimates, to the psychiatrist.

Table 2 shows the result of separating the cases into two categories, 'anxious' or 'depressed' (defined to be mutually exclusive), and the eight categories are defined by the same three factors as before. Does the particular diagnosis of a case (i.e. anxious rather than depressed) depend at all on any of the factors?

Table 1

Case	Total	rmq	mcr	events
19	81	1	1	1
5	9	2	1	1
1	4	1	2	1
4	4	2	2	1
38	66	1	1	2
14	15	2	1	2
12	13	1	2	2
7	8	2	2	2

Table 2

anx	dep	Case	rmq	mcr	events
13	6	19	1	1	1
5	0	5	2	1	1
0	1	1	1	2	1
2	2	4	2	2	1
28	10	38	1	1	2
7	7	14	2	1	2
9	3	12	1	2	2
4	3	7	2	2	2

[This illustrates the use of binomial regression.]

9. Agresti (1990) *Categorical Data Analysis*, p. 377, gives the table below, relating mother's education to father's education for a sample of eminent black Americans (defined as persons having a biographical sketch in the publication *Who's Who Among Black Americans*).

Mother's	Father's education			
education	1	2	3	4
1	81	3	9	11
2	14	8	9	6
3	43	7	43	18
4	21	6	24	87

Here, for education

1 = 8th grade or less
2 = Part High School
3 = High School
4 = College.

Let $p_{ij} = Pr$ (mother's education is i, father's education is j), $1 \leq i, j \leq 4$. Consider the model

$$\omega : \left.\begin{array}{ll} p_{ij} = \theta\varphi_i + (1-\theta)\alpha_i\beta_j & \text{for } i = j \\ p_{ij} = \qquad\quad (1-\theta)\alpha_i\beta_j & \text{for } i \neq j \end{array}\right\}$$

where $\Sigma\varphi_i = 1$, $\Sigma\alpha_i = 1$, $\Sigma\beta_j = 1$, and $0 < \theta < 1$. Can you interpret ω to a sociologist?

Show that, under ω,

$$\log p_{ij} = a_i + b_j \quad \text{for } i \neq j.$$

With n, the cell frequencies, declared as Poisson variables, with the default link function, and with the two factors M.Ed and F.Ed each with the 4 given values, glm finds that
$n \sim M.Ed + F.Ed$ has deviance 159.25, with 9 df.
Why should you expect this deviance to be so large?
But if we omit the 4 diagonal entries of the table, and fit
$n \sim M.Ed + F.Ed$, we find that the deviance is 4.6199, with 5 df. How do you interpret this?
Find the fitted frequencies for this latter model.

10. You see below the results of using glm to analyse data from Agresti(1996, p247) on tennis matches between 5 top women tennis players (1989-90). Let (r_{ij}) be the number of wins of player i against player j, and let n_{ij} be the total number of matches of i against j, for $1 \leq i < j \leq 5$. (Thus we have 10 observations, which we will assume are independent binomial, with $E(r_{ij}) = n_{ij}p_{ij}$.)
The model we will fit is

$$\log(p_{ij}/(1-p_{ij})) = \alpha_i - \alpha_j, \quad \text{with } \alpha_5 = 0 \quad i < j$$

equivalently

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, \quad \text{where } \pi_i = e^{\alpha_i}.$$

The data can be read in (read.table(" ", header=T)) as the table

```
wins tot sel graf saba navr sanc
2 5 1 -1 0 0 0
1 1 1 0 -1 0 0
3 6 1 0 0 -1 0
2 2 1 0 0 0 -1
6 9 0 1 -1 0 0
3 3 0 1 0 -1 0
7 8 0 1 0 0 -1
1 3 0 0 1 -1 0
```

3 5 0 0 1 0 -1
3 4 0 0 0 1 -1

Thus for example, the first row of numbers tells us that 'sel' played 'graf' for a total of 5 matches, and 'sel' won 2 of these. The result of
$glm(wins/tot \sim sel + graf + saba + navr - 1, binomial, weights = tot)$
is deviance 4.6493, with 6 df.
The parameter estimates are
$sel = 1.533(0.7870), graf = 1.933(0.6783), saba = 0.731(0.6770), navr = 1.087(0.7236)$, with $sanc = 0$ by assumption.

(i) Why do we impose a constraint on $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$?

(ii) Can we confidently say that Graf is better than Sanchez?

(iii) Can we confidently say that Graf is better than Seles?

(iv) What is your estimate of the probability that Sabatini beats Sanchez, in a single match? (Answer: 0.6750)

(v) Here are the corresponding data for 5 top men tennis players during 1989–90. Analyse them, fitting a model of the same form as above.

	Loser				
Winner	Edberg	Lendl	Agassi	Sampras	Becker
Edberg	–	5	3	2	4
Lendl	4	–	3	1	2
Agassi	2	0	–	1	3
Sampras	0	1	2	–	0
Becker	6	4	2	1	–

11. The purpose of this example is to introduce you to the topic of *overdispersion* in the context of the Poisson distribution.

Suppose now that the observations $Y_1, ..., Y_n$ are independent, with $E(Y_i) = \mu_i$, and $var(Y_i) = \phi\mu_i$, and $log(\mu_i) = \beta x_i$, for some unknown ϕ and unknown scalar parameter β. Let β_0 be the true value of this unknown parameter.

Our aim is to estimate β, but ϕ is an unknown 'dispersion' parameter. Clearly $\phi > 1$ will correspond to over-dispersion relative to the Poisson. In the absence of knowledge of ϕ, we choose our estimator $\hat{\beta}$ to maximise the function $l_p(\beta)$, where

$$l_p(\beta) = -\Sigma\mu_i + \beta\Sigma x_i y_i + \text{constant}.$$

(Thus $l_p()$ is in general not the 'correct' loglikelihood function: we work out below whether this is a serious problem.)

By expanding

$$\frac{\partial l_p(\beta)}{\partial\beta}$$

evaluated at $\hat{\beta}$, about β_0, show that $(\hat{\beta} - \beta_0)$ is approximately equal to $(I(\beta_0))^{-1}U(\beta_0)$, where

$$U(\beta) = \frac{\partial l_p(\beta)}{\partial\beta}, \text{ and } I(\beta) = \Sigma x_i^2 \exp\beta x_i,$$

and hence show that, approximately,

$$E(\hat{\beta}) = \beta_0, \text{ and } var(\hat{\beta}) = \phi(I(\beta_0))^{-1}.$$

Thus if $\phi > 1$, the true variance of the $\hat{\beta}$ will be greater than the value given by software which assumes the Poisson distribution.

R allows you to make this simple correction for overdispersion, by

summary(glm(y ~ x, poisson), dispersion =3.21))

where, for example, the 'dispersion' ϕ has been estimated as (deviance/df), here 3.21. Experiment with this.

Now generalise your result to the case where β, x_i are vectors of dimension p.