UNIVERSITY OF LONDON

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)
MAY–JUNE 2003

This paper is also taken for the relevant examination for the Associateship.

M3S12 BIOSTATISTICS

DATE: Wednesday, 28th May 2003 TIME: 10 am – 12 noon

Credit will be given for all questions attempted but extra credit will be given for complete or nearly complete answers.

Calculators may not be used. Statistical tables will not be available.

1. *a)* In an epidemiological study, *exposure* is a binary risk factor, and individuals in the study are categorized as *exposed* or *not exposed* to the risk factor, denoted E and E' respectively. Disease *incidence* for an individual is denoted F (*affected*) and its complement F' (*not affected*).

Identify the principal difference, in terms of exposure, incidence and inclusion in the study between

i) *observational* and *experimental* epidemiological studies,

ii) *cohort* and *case-control* studies.

In terms of the events E and F, and conditional probability notation, define the following *measures of effect*; in each case, state whether the quantity is estimable from a *cohort* study and a *case-control* study - where appropriate, give the form of the estimate of the quantity derived from a sample of data cross-categorized in the usual 2×2 table fashion.

iii) the *incidence probability* in the exposed group,

iv) the *relative risk* of disease in the exposed/unexposed groups,

v) the *odds-ratio*.

b) Data from a cohort study involving the risk factor age and its impact on a particular psychiatric disorder for a particular population are available. There are five age categories: for each category, let D denote the number of deaths, and N denote the total number of person-years on study.

AGE GROUP	D	N
10-19	20	4000
20-29	150	6000
30-39	120	4000
40-49	80	4000
50+	10	2000

i) Compute and report in an appropriate form the *crude incidence rate* of the disorder.

i) Explain and illustrate the difference between the *crude*, *specific*, and *standardized* incidence rates in this context.

ii) Give an expression for the *standardized* incidence rate for a hypothetical standardizing population for which the breakdown across the five age categories is $(25\%, 30\%, 25\%, 10\%, 10\%)$.

2. In a small cohort study of patients who have undergone cruciate ligament reconstruction surgery, the effectiveness of two types of operation are to be compared. Here, exposure E corresponds to operation type I (patella graft), and E' corresponds to operation type II (hamstring graft); disease incidence F corresponds to the failure of the reconstruction within two years of the surgery. The data can be denoted in the usual cross-categorized 2×2 table fashion (exposure status in the columns, health status in the rows) as $(n_{11}, n_{12}, n_{21}, n_{22})$, with row totals $(n_{1\cdot}, n_{2\cdot})$ and column totals $(n_{\cdot 1}, n_{\cdot 2})$; the data available are $n_{11} = 44, n_{12} = 26, n_{21} = 1002, n_{22} = 247$. Denote by γ_1 and γ_0 the *exposure rates* in the disease (case) and healthy (control) groups respectively. Throughout this question, the notation log refers to **natural logarithm**.

i) Show that the maximum likelihood estimate of γ_1 is $\widehat{\gamma}_1 = n_{11}/n_{1\cdot}$. State the asymptotic normal distribution of the corresponding maximum likelihood estimator, and the form of the *estimated standard error* for $\widehat{\gamma}_1$.

ii) The *relative exposure rate* is $\tau = \gamma_1/\gamma_0$. Find the maximum likelihood estimate of τ, $\widehat{\tau}$, and show that the estimated standard error for $\log \tau$ is

$$s.e.\left(\log \widehat{\tau}\right) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1\cdot}} + \frac{1}{n_{21}} - \frac{1}{n_{2\cdot}}}.$$

Use the result that for random variables U_n and $V_n = g(U_n)$ for differentiable function g,

$$U_n \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ then } V_n \dot{\sim} N\left(g(\mu), \frac{\sigma^2 \{g'(\mu)\}^2}{n}\right).$$

iii) State (without proof) the form of the estimated standard error for the *log odds ratio*, $\log \psi$.

iv) Compute an *approximate 95% confidence interval* for the log odds ratio using the following numerical results:

$$\log \widehat{\psi} = -0.8743, \qquad s.e.\left(\log \widehat{\psi}\right) = 0.2574.$$

Hence assess the evidence for a difference in the outcome for the two types of surgery.

v) Derive the Bayesian posterior distribution of γ_1 if the prior distribution is a $Beta\,(\alpha_1, \beta_1)$ distribution.

3. *a)* Binary incidence data, y_i for $i = 1, ..., n$, are to be collected, where $y_i = 1$ indicates that individual i was a sufferer from the disease concerned. The dependence of the incidence probability on a predictor is to be studied using a particular Generalized Linear Model.

i) Describe the key aspects of a *logistic regression model* for the individual level data. Outline methods for hypothesis testing of the importance of the predictor.

ii) Suppose that data for a single continuous predictor is recorded. Give details of the *linear predictors* for the three principal models that may be fitted to the response y, namely the *null*, *main effect* and *saturated* models.

iii) Derive the form of the *deviance* for the main effect model logistic regression for response y. Outline how deviance can be used to assess and compare the fit of a GLM.

b) A method of predicting whether a pregnant woman will require a Caesarian section ($y = 1$) as opposed to a natural birth ($y = 0$) is required. Body mass index (BMI), that is weight/(height)^2, at the beginning of pregnancy is thought to be a good predictor for the eventual childbirth method.

In a study, the childbirth method for $n = 920$ women was recorded, along with their initial BMI, which was discretized into the four categories, $[0, 20)$, $[20, 30)$, $[30, 40)$ and $[40, \infty)$ (units kg/m^2). A summary of the data is presented below; s_i is the total number of women who had a Caesarian section, and n_i is the number of women in the ith BMI subgroup.

BMI (kg/m^2)	$[0, 20)$	$[20, 30)$	$[30, 40)$	$[40, \infty)$
s_i	49	97	14	1
n_i	425	450	43	2

An SPLUS analysis of these data is given in OUTPUT 1 (page 7); the treatment-contrasts parameterization is used; the `Coefficients` output are baseline (`Intercept`) and differences from baseline on the linear predictor scale.

i) Is BMI a useful predictor of childbirth method ? Justify your answer.

ii) Comment on the fit of the main effect model.

ii) Would the analysis be improved if individual-level BMI data was retained, so that BMI could be included in the model as a continuous predictor ? Justify your answer.

4. A Poisson model is deemed relevant for the incidence data below.

Count y	Person-years d	Exposure E	Age group A	Age category
1	15382.27	0	0	$[0, 25)$
6	19946.65	1	0	$[0, 25)$
0	31413.31	0	1	$[25, 40)$
16	26503.54	1	1	$[25, 40)$
1	33727.60	0	2	$[40, 60)$
9	16407.93	1	2	$[40, 60)$
2	38069.95	0	3	60^+
7	36492.51	1	3	60^+

The counts y are the numbers of cases of a rare form of cancer, the person-years data d relate (approximately) to the total time on study of a cohort accumulated over a number of years, exposure E has two levels, with level 1 indicating close proximity (within 2km) to a commercial incinerator site, and age group A is a potential confounder having four levels. The expected value of Y is thought to depend linearly on d.

a) *i)* Write down a generalized linear model (GLM) appropriate for the analysis of these data. Explain the importance of an *offset* term in the model.

ii) List the number of parameters that each of the following models (defined in standard notation) contains:

$$NULL, E, A, E + A, E * A$$

(you may write down the numbers without further justification). Identify the *saturated* model.

iii) Derive the general form of the *deviance residual* for this model.

b) An SPLUS analysis of deviance of these data is summarized on pages 8 and 9.

i) Find the most appropriate model (in terms of deviance) for the data. Justify your conclusion.

ii) Is there any evidence of *overdispersion* in the data (relative fit of your preferred model) ? Justify your answer.

iii) Briefly outline the *quasilikelihood* approach to modelling overdispersed data.

5. *a)* Suppose that $\{X_{ij} : i = 1, 2 \text{ and } j = 1, 2\}$ are independent Poisson random variables with parameters $\{\lambda_{ij} : i = 1, 2 \text{ and } j = 1, 2\}$ respectively.

i) Consider the new random variables $\{Y_1, Y_2, Y_3, Y_4\}$ where

$$Y_1 = X_{11} \qquad Y_2 = X_{12} \qquad Y_3 = X_{21} \qquad Y_4 = X_{11} + X_{12} + X_{21} + X_{22}$$

Find the joint conditional mass function of (Y_1, Y_2, Y_3) given that $Y_4 = n_{..}$ for some $n_{..} > 0$, and explain the relevance of this result for fitting models to contingency table data.

ii) Suppose that a model that presumes *symmetry* in the 2×2 table of Xs, that, is, that

$$\lambda_{12} = \lambda_{21} = \lambda,$$

say, is to be considered. Derive the maximum likelihood estimates of the three parameters in the model $(\lambda_{11}, \lambda_{22}, \lambda)$.

b) The *Pearson Chi-Squared* goodness-of-fit statistic for a general contingency table *containing Poisson data* with cell entries $\{n_{ij}, i = 1, ..., I, j = 1, ..., J\}$ takes the form

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

where $\widehat{n}_{ij}$ is the fitted cell entry for cell (i, j) under the model, and where, asymptotically, if the fitted model is adequate

$$X^2 \ \dot{\sim}\ \chi^2_{IJ-d}$$

where, here, d is the number of parameters estimated in the model.

i) Find the form of X^2 and its asymptotic distribution for the symmetry model in *a) ii)*.

ii) Use this X^2 statistic to test the symmetry model in the following 2×2 table; the data concerned relate to the health status (Prone to Colds/Not Prone to colds) at the beginning of the trial and the end of the follow up, for each of 200 school children.

		End of Study	
		Not Prone	Prone
Start of Study	Not Prone	160	10
	Prone	20	10

OUTPUT 1 FOR QUESTION 3

```
> summary(glm(Y ~ factor(BMI.GROUP), family = binomial, data = bmi.data))

Call: glm(formula = Y ~ factor(BMI.GROUP), family = binomial, data = bmi.data)

Coefficients:
                        Value  Std. Error  t value
       (Intercept) -2.0377670  0.1517525 -13.428228
factor(BMI.GROUP)2  0.7460099  0.1901864   3.922519
factor(BMI.GROUP)3  1.3095285  0.3590824   3.646875
factor(BMI.GROUP)4  2.0377670  1.4223322   1.432694

    Null Deviance: 853.2566 on 919 degrees of freedom
Residual Deviance: 829.9679 on 916 degrees of freedom

------------------------------------------------------------------------

Table: 0.95 quantiles of Chisquared(DF) distribution

           DF  Quantile
            1    3.8415
            2    5.9915
            3    7.8147
            4    9.4877
            5   11.0705
          915  986.4829
          916  987.5214
          917  988.5598
          918  989.5982
          919  990.6366
          920  991.6750

-------------------------------------------------------------------------
```

DEVIANCE SUMMARY FOR QUESTION 4

MODEL	DF	D	$\triangle DF$	$\triangle D$	$\chi^2_{\triangle DF}(0.95)$
$NULL$	7	49.53132	-	-	-
E	6	11.37614	1	38.15518	3.8415
A	4	45.35049	3	4.18083	7.8147
$E+A$	3	5.58927	4	43.94205	9.4877
$E*A$	0	0.000	7	49.53132	14.0671

OUTPUT 2 FOR QUESTION 4

```
NULL MODEL
----------
               Value   Std.Error  t value
(Intercept) -8.554285  0.1536309 -55.68076

    Null Deviance: 49.53132 on 7 degrees of freedom
Residual Deviance: 49.53132 on 7 degrees of freedom

----------------------------------------------------------------------------

MAIN EFFECT MODEL E
-------------------

                 Value  Std.Error    t value
(Intercept) -10.297159  0.4999313 -20.597149
  factor(E)   2.428335  0.5255917   4.620193

    Null Deviance: 49.53132 on 7 degrees of freedom
Residual Deviance: 11.37614 on 6 degrees of freedom

-----------------------------------------------------------------------------

MAIN EFFECT MODEL A
-------------------


                  Value  Std.Error    t value
(Intercept) -8.52654725  0.3779645 -22.55912373
 factor(A)1  0.33237227  0.4531495   0.73347158
 factor(A)2  0.00664716  0.4927853   0.01348896
 factor(A)3 -0.49562051  0.5039526  -0.98346647

    Null Deviance: 49.53132 on 7 degrees of freedom
Residual Deviance: 45.35049 on 4 degrees of freedom

-----------------------------------------------------------------------------

MAIN EFFECTS MODEL E+A
----------------------

                  Value  Std.Error    t value
(Intercept) -10.5133734  0.6229572 -16.8765590
  factor(E)   2.4969047  0.5268842   4.7390007
 factor(A)1   0.5109059  0.4533307   1.1270049
 factor(A)2   0.4571701  0.4947232   0.9240927
 factor(A)3  -0.3735754  0.5039551  -0.7412872

     Null Deviance: 49.53132 on 7 degrees of freedom
 Residual Deviance: 5.558927 on 3 degrees of freedom
```