Imperial College
London

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2008

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

M3S12/M4S12

Biostatistics

Date: Monday, 2nd June 2008 Time: 2 pm – 4 pm

Answer all questions. Each question carries equal weight.

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.

1. Suppose that we have observed data x, which is assumed to be distributed according to a probability density function (pdf) $f(x|\theta)$ for $\theta \in \mathbb{R}^k$, and that prior to observing the data the parameter θ is assigned a prior probability density $p(\theta)$.

(i) Derive, from the axioms of probability, Bayes theorem:

$$\pi(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int_{\mathbb{R}^k} f(x|\theta)p(\theta)d\theta}$$

and explain its interpretation in relation to the data and parameter.

Now suppose that the data are comprised of a response variable Y and predictor variable X, that is, the data are $(y_1, x_1, \ldots, y_n, x_n)$.

(ii) (a) Show that the likelihood for the normal linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $\beta_0, \beta_1 \in \mathbb{R}$, $i = 1, \ldots, n$, and ϵ_i are independently distributed $N(0, \sigma^2)$ is:

$$f_{Y|\beta,\sigma^2}(y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) \right\}$$

where $\mathbf{X}$ is the design matrix.
(b) State the assumptions of the model; what are the implications, in terms of the data?

(iii) Suppose that an investigator measures the blood pressure Y_i (in millibars Hg) and the age X_i (in years) of n individuals. It is hypothesised that the relationship between the age and blood pressure is constant below some known age, and then grows linearly. That is to say, the model

$$Y_i = \begin{cases} \beta_0 + \epsilon_i & \text{if } X_i \leq c \\ \beta_1 + \beta_2 X_i + \epsilon_i & \text{if } X_i > c \end{cases}$$

is proposed, with ϵ_i as above.
(a) Rewrite the model in the form:

$$Y_i = \sum_{j=0}^{2} \beta_j B_j(X_i) + \epsilon_i$$

for some B_0, B_1 and B_2 to be found.
(b) Suppose that we assume that, prior to observing the data, we have that

$$p(\beta, \sigma^2) = \frac{1}{(2\pi)^{1/2}} \frac{1}{(\sigma^2)^{3/2}} \exp\left\{ -\frac{1}{2\sigma^2}\beta^T\beta \right\} \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left\{ -\frac{b}{\sigma^2} \right\}$$

with $a, b > 0$. Derive the joint posterior distribution $p_{\beta,\sigma^2|X,Y}(\beta, \sigma^2|X, Y)$.
(c) Discuss how this model could be compared, formally, with the normal linear model in (ii) (a).

2. (i) (a) What is a generalized linear model (GLM)? What is the central idea behind its construction?
(b) Define the linear predictor and link function.
(c) What is the logistic regression model? In this case, show that the probability density function of the response belongs to the exponential family.
(d) What is the deviance of a GLM? What is it used for?

(ii) The institute of the Child Health Policy at the University of Florida studies the effects of health policy decisions on children's health. An example of such a study is as follows. The overall health of a child is described by two quantitative variables A and B. Each child is in an *Health Maintenance Organization* (HMO), which is either profit or non-profit $X \in \{-1, 1\}$. The study is interested in a response variable $Y_{ij} \in \{0, 1\}$, ($i = 1, 2$, $j = 1, \ldots, n_i$) of whether an emergency room is used or not, depending upon the status of the HMO and the child's health. The data that are observed correspond to n_1 individuals that are in non-profit HMOs and n_2 in profit HMOs.
(a) Suppose that we model the logit of probability of using an emergency room, p_{ij}. Derive the likelihood of an appropriate GLM with linear predictor that is additive in each of the explanatory variables.
(b) It is assumed that X is always included in the GLM in (a). We are given the following output, for our observed data:

Model	DF	D	ΔDF	ΔD	$\chi^2_{\Delta DF}(0.95)$
X	98	48.73	–	–	–
$A + X$	97	11.57	1	37.16	3.84
$B + X$	97	47.52	1	1.21	3.84
$A + B + X$	96	5.53	2	43.20	5.99

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

3. (i) A two way $I \times J$ contingency table of count data is assumed to have entries n_{ij} that are realizations of independent random variables N_{ij} that are Poisson distributed:

$$N_{ij} \sim Poisson(\lambda_{ij})$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$. Find the probability distribution of the entries in the table conditional on the row totals $(N_{1.}, \ldots, N_{I.})$ taking their observed values $(n_{1.}, \ldots, n_{I.})$, where

$$n_{i.} = \sum_{j=1}^{J} n_{ij}.$$

(ii) The following data are from a small cohort study carried out to discover the relationship between the exposure to radiation of the fathers of children employed at a nuclear power plant, and incidences of leukemia, collated over a ten year period

	Exposed	Non-Exposed
Leukemia	11	3
Healthy	19	37

Is there any statistical evidence of association or dependence between the row and column factors? Justify your answer, commenting on the validity of the chosen statistical test (note $\chi^2_1(0.95) = 3.841$).

Explain briefly how and why conditioning might be used in the analysis of the data in this table.

4. (i) 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

(a) observational and experimental epidemiological studies

(b) 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.

(c) the incidence probability in the exposed group,

(d) the relative risk of disease in the exposed/unexposed groups,

(e) the odds-ratio.

(ii) 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
0-19	20	4000
20-29	150	6000
30-39	120	4000
40-49	80	4000
50+	10	2000

(a) Compute and report in an appropriate form the crude incidence rate of the disorder.

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

(c) 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%).

5. Suppose that, in an inference setting, a decision is to be made from a collection of alternatives $d \in \mathcal{D}$. We regard a parameter, $\theta \in \mathbb{R}$, in a statistical model which can be associated with such a decision.

(i) (a) Define a loss function.

(b) What is the Bayes rule?

(c) Derive the Bayes rule for estimating θ when a quadratic loss function is adopted. You may assume that all relevant expectations are finite.

(ii) Government agencies use mortality forecasts when planning and developing health policy. These agencies rely upon mortality predictions in deciding upon the allocation of funds from government sources. In a recent study, the logarithm of the mortality rate at an observed time, $1 \leq n \leq p$, $Y_n \in \mathbb{R}$ is thought to be related to an unobserved random variable $X_n \in \mathbb{R}$, through the following relationship

$$\begin{aligned} Y_n &= X_n + \epsilon_n \\ X_n &= X_{n-1} + \nu_n \end{aligned}$$

where, $X_0 = 0$, for any $n \geq 1$, ϵ_n are independent $N(0, \sigma_\epsilon^2)$ random variables and, independently of ϵ_n, ν_n are independent $N(0, \sigma_\nu^2)$; both σ_ϵ^2 and σ_ν^2 are known parameters.

(a) You are **given** that

$$X_1 | y_1 \sim N(\mu_1, \sigma_1^2)$$

where

$$\sigma_1^2 = \frac{\sigma_\epsilon^2 \sigma_\nu^2}{\sigma_\epsilon^2 + \sigma_\nu^2} \quad \text{and} \quad \mu_1 = \frac{\sigma_1^2 y_1}{\sigma_\epsilon^2}.$$

Hence, complete a proof by induction, that the posterior distribution, for any $n \geq 1$, $\pi(x_n | y_1, \ldots, y_n)$, is normal with mean μ_n and variance σ_n^2, with explicit recursions for μ_n and σ_n^2.

(b) Find the Bayes estimate, under squared error loss, of X_{p+1}, given only the observed data. Discuss how this estimate could be used to forecast Y_{p+1}.