UNIVERSITY OF LONDON

GOLDSMITHS' COLLEGE

B. Sc. Examination 2003

STATISTICS

ST52011A (ST230) Statistical Modelling

Duration: 3 hours

Date and time:

Answer four questions. You must answer at least ONE question in each section. There are 100 marks available on this paper. Each question is worth 25 marks.

Electronic calculators may be used. The make and model should be specified on the script. The calculator must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

NOTE: Full details of all calculations are to be shown; pre-programmed statistical tests and procedures on a calculator, apart from mean and standard deviation, must not be used.

WHITE, YEATS & SKIPWORTH: Tables for Statisticians to be provided.

A formula sheet is attached at the end of this paper.

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Section A

Question 1 You must answer this question.

(a) An engineer is studying the rate of metal removal in a machining operation. The rate of removal is related to the cutting speed and the hardness of the test specimen. She first takes eight observations at a cutting speed of 1000 revolutions per minute which are shown in the following table

y	x
68	120
90	140
98	150
77	125
88	136
110	175
65	110
79	130

where y is the amount of metal removed and x is the hardness of the specimen. She enters these data into MINITAB and some of the output produced is shown below.

(i) The engineer assumes the data follow a simple linear regression model. Write down this model and the assumptions made. [2]

(ii) What conclusions do you draw from the Analysis of Variance table? [1]

(iii) A plot of standardised residuals versus fitted values shows a fairly random scatter. What does this imply? Sketch such a plot which might cause you to doubt one of the assumptions. [2]

(iv) Comment on the normal plot of the standardised residuals. [2]

(v) Define leverage. Comment on the plot of leverage versus i. [2]

(vi) Define Cook's distance and comment on the plot of it versus i. [3]

```
Regression Analysis: y versus x

The regression equation is
y = -19.4758 + 0.760411 x

S = 3.60643      R-Sq = 95.4 %      R-Sq(adj) = 94.7 %
```

```
Analysis of Variance

Source            DF         SS         MS         F      P
Regression         1    1631.46    1631.46   125.436  0.000
Error              6      78.04      13.01
Total              7    1709.50

Inverse Cumulative Distribution Function

F distribution with 2 DF in numerator and 6 DF in denominator

P( X <= x )          x
    0.5000        0.7798
```

plot for st230q1(a)

plots for st230q1(a)

(b) The engineer goes on to collect data at cutting speeds of 1200rpm and 1400rpm which are shown below.

1200rpm		1400rpm	
y	x	y	x
112	165	118	175
94	140	82	132
65	120	73	124
74	125	92	141
85	133	80	130

She further assumes that simple linear regression model holds at each cutting speed and asks her assistant to plot the data and calculate the necessary sums of squares and products not only for each cutting speed group, but also for the overall data set of 18 observations. His results are in the following table:

	1000rpm	1200rpm	1400rpm	overall
$CS(y,y)$	1709.5	1326.0	1236.0	4356.50
$CS(x,y)$	2146.5	1257.0	1420.0	$4896.8\dot{3}$
$CS(x,x)$	2821.5	1241.2	1645.2	$5777.6\dot{1}$
RSS	78.038	52.999	10.374	206.173

The engineer requires to test the hypothesis that her three simple linear regression models have parallel slopes.

(i) What are the full and reduced models for the hypothesis of parallelism? [1]

(ii) Find the least squares estimate of the common slope under the reduced model. (You may quote suitable formulae without proof). [3]

(iii) Show that there is no evidence against parallelism by obtaining the RSS of both models and performing an appropriate F-test. [5]

(iv) Show further that there is also no evidence against the hypothesis that her three simple linear regression models have also the same intercept. [3]

(v) What should the engineer conclude? [1]

Section B

You must answer at least one question from this section.

Question 2 A study in the 1970's in the US considered the variables which were important in determining the selling price of a house. The following independent variables were considered:

taxes	Property taxes (hundreds of dollars)
baths	Number of baths
area	Area of plot (thousands of square feet)
space	Living space (thousands of square feet)
garages	Number of garages
rooms	Number of rooms
bedrooms	Number of bedrooms
age	Age of home (years)
firepl	Number of fireplaces

The dependent variable (price) was the selling price in thousands of dollars. Data were collected on 28 houses.

(a) A statistician entered the data into MINITAB and produced the following output.

```
Stepwise Regression: sales versus taxes, baths, ...

Forward selection.  Alpha-to-Enter: 0.05

 Response is  sales   on  9 predictors, with N =   28


    Step          1        2        3
Constant    -1.4669  -0.6315  -1.0658

baths          31.3     17.8      8.2
T-Value       12.42     4.05     1.71
P-Value       0.000    0.000    0.100

taxes                   2.26     1.90
T-Value                 3.50     3.37
P-Value                0.002    0.003
```

```
space                          10.1
T-Value                        3.22
P-Value                       0.004

S             5.48     4.58     3.90
R-Sq         85.58    90.32    93.24
R-Sq(adj)    85.02    89.54    92.39
C-p           21.6      8.6      1.4
```

(i) Explain what fitting method is being used. [2]

(ii) Write down the model selected by this method. [1]

(iii) Comment on the fact that in the final model the coefficient for baths has a P-value of 0.100. [3]

(iv) Define the quantity R-Sq given in the output. Comment on its use to aid model choice. [3]

(v) Define the quantity R-Sq(adj) given in the output. In what way is it an improvement over R-Sq? [3]

(vi) Define the quantity C-p given in the output. Explain how it is used to aid model choice. [4]

(b) Explain the method of backwards fitting. What is a possible problem in fitting the full model with many variables? [6]

(c) Could the method of backwards fitting result in the same model as selected in part (a)? Justify your answer. [3]

Question 3 In an experiment to compare three different varieties of corn 15 plots in a field were allocated one of the three varieties at random. The yields (in bushels per plot) of the three varieties were as follows

Corn variety		
X	Y	Z
22.6	27.6	25.4
21.5	26.5	24.5
22.1	27.0	26.3
21.8	27.2	24.8
22.4	26.8	25.1

(a) Write down a suitable model for these data and any necessary assumptions. [4]

(b) Draw up the analysis of variance table and test for differences between the varieties. [7]

(c) Describe how to check the model assumptions using MINITAB

(i) graphically, [2]

(ii) using appropriate tests. [2]

(d) Y and Z are known to be related varieties and X is unrelated to the other two. Suggest appropriate mutually orthogonal contrasts for testing, explaining what hypothesis each of your contrasts tests, and carry out these tests. [10]

Question 4 A marketing organisation wishes to study the effects of four sales methods on weekly sales of a product. The organisation uses a randomised block design in which three salesmen use each of the sales methods. The results obtained are given below.

		Salesman		
		A	B	C
Method	1	32	29	30
	2	32	30	28
	3	28	25	23
	4	25	24	23

(a) Write down the model used and the assumptions made. [4]

(b) What is the advantage of using a randomised block design here rather than a completely randomised design? [4]

(c) Carry out a graphical analysis that shows that little interaction exists between sales method and salesmen. [6]

(d) Draw up the analysis of variance table and test for differences between the (i) sales methods and (ii) the salesmen. [11]

Section C

You must answer at least one question from this section.

Question 5 Accidents at a busy junction are thought to follow a Poisson process so that the number of accidents X_i in a time period of length t_i (known) has a Poisson distribution with mean $\mu_i = \lambda t_i$ $(i = 1, \ldots, n)$, where $\lambda > 0$ is an unknown parameter. The parameter λ represents the accident rate that is to be estimated from the n non-overlapping periods of observation.

(a) Show that the likelihood function $L(\lambda)$ is proportional to $e^{\lambda \sum_{i=1}^{n} t_i} \lambda^{\sum_{i=1}^{n} x_i}$. [4]

(b) Sketch the likelihood function. [2]

(c) Show that the maximum likelihood estimate (MLE) $\hat{\lambda}$ satisfies

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} t_i}.$$ [3]

(d) Show that an approximate 95% confidence interval for λ is given by the end-points $(\sum_{i=1}^{n} x_i \pm 1.96\sqrt{\sum_{i=1}^{n} x_i})/\sum_{i=1}^{n} t_i$. [6]

(e) The *saturated model* postulates a different accident rate λ_i, say for the i^{th} period $(i = 1, \ldots, n)$ and hence the MLEs are $\dot{\lambda}_i = x_i/t_i$, $i = 1, \ldots, n$. Show that the deviance of the null model above is given by the expression

$$2\sum_{i=1}^{n} x_i \left[\log\left(\frac{x_i}{\sum_{i=1}^{n} x_i}\right) - \log\left(\frac{t_i}{\sum_{i=1}^{n} t_i}\right)\right].$$ [10]

Question 6 A survey of women gave the proportions who have reached the menopause in each of five 2-year age groups which were further subdivided into smokers and non-smokers. The results and some MINITAB output are given below:

Numbers of menopausal women by age and smoking habit

Age (years)	Total Smokers	Number Menopausal	Total Non-Smokers	Number Menopausal
45	67	1	37	1
47	44	5	29	5
49	66	15	36	13
51	73	33	43	26
53	52	37	28	25

Plots of observed proportions and logits by age and smoking (1/+) or not (0)

Binary Logistic Regression: number, total versus smoke, age

Link Function: Logit

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-30.826	4.641	-6.64	0.000			
smoke							
1	2.716	5.864	0.46	0.643	15.11	0.00	1.48E+06
age	0.61629	0.09317	6.61	0.000	1.85	1.54	2.22
smoke*age							
1	-0.0684	0.1172	-0.58	0.559	0.93	0.74	1.17

Log-Likelihood = -218.663
Test that all slopes are zero: G = 170.993, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	2.180	6	0.902
Deviance	2.291	6	0.891

Binary Logistic Regression: number, total versus smoke, age

Link Function: Logit

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-28.743	2.816	-10.21	0.000			
smoke							
1	-0.7068	0.2463	-2.87	0.004	0.49	0.30	0.80
age	0.57441	0.05651	10.17	0.000	1.78	1.59	1.98

Log-Likelihood = -218.836
Test that all slopes are zero: G = 170.647, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	2.512	7	0.926
Deviance	2.637	7	0.916

```
Matrix XPWX3

 7.93204  -0.15878   0.07677
-0.15878   0.00319  -0.00228
 0.07677  -0.00228   0.06066
```

(a) Describe fully what models are being fitted and perform simple goodness of fit tests for each (in the absence of suitable residual plots) [10]

(b) Perform a suitable test of the hypothesis that the effect of age on the probability of the menopause is the same for both smokers and non-smokers (on the logit scale). [4]

(c) Estimate the effect of smoking as an equivalent increase in menopausal age, and derive an approximate 95% confidence interval for this increase, using the formula below with suitable estimates inserted from the MINITAB output

$$\mathrm{Var}\left[\frac{U}{V}\right] \approx \frac{1}{(E[V])^2}\left[\mathrm{Var}[U] - 2\frac{E[U]}{E[V]}\mathrm{Cov}[U,V] + \left(\frac{E[U]}{E[V]}\right)^2 \mathrm{Var}[V]\right]. \qquad [11]$$

Question 7 A random sample of n students perform an experiment in which the i^{th} student tosses a drawing pin k_i times and notes the number of times x_i that the pin falls uppermost ($i = 1, \ldots, n$).

(a) Find the likelihood for the unknown propensity p of the pin to fall uppermost, and sketch it, ignoring any proportionality constants. [5]

(b) Hence find the maximum likelihood estimate $\hat{p}$, and the information function $I(p)$ which gives the classical 95% confidence interval $\hat{p} \pm 1.96/\sqrt{I(\hat{p})}$. [7]

(c) What is a suitable family of prior distributions for a Bayesian to choose from? [1]

(d) On the assumption of an arbitrary choice from this family, find the posterior mean $E[p \mid \text{data}]$ and show it is a weighted linear combination of prior mean and MLE. [4]

(e) Compare the symmetric Bayesian interval centred on the posterior mean and using a normal approximation with that of (b), under a suitably chosen prior representing great uncertainty (GPU). [8]