UNIVERSITY OF CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

9am Tuesday 13 June to 1pm Friday 16 June 2006

APPLIED STATISTICS

*Attempt **THREE** questions. There are **FOUR** questions in total.*
Marks for each question are indicated on the paper in square brackets.
Each question is worth a total of 20 marks.

This is an 'Open Book' examination, involving the use of the Statistical Laboratory's network of workstations. Candidates will receive this paper at 9.00am on Tuesday 13 June, and must hand in their scripts to the Chairman of Examiners by 1.00pm on Fridy 16 June.

The data sets will be emailed to candidates on Tuesday 13 June.

(The Statistical Laboratory Computer Officer and Examiner will normally be available for consultation if required between 9.00am and 4.30pm on these four days.)

Each candidate should submit his/her script with a signed statement that the work has been carried out without any collaboration with others.

The scripts may be handwritten. Candidates are requested to submit at most 25 pages in total. They are advised that the total work set should take between 4 and 6 hours.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 The table below appeared in the Times (17 October 2005) under the heading "Caught on camera: the effect of speeding convictions upon insurance premiums (in £)".

	Number of points			
	0	3	6	9
21 year old male	306	384	384	409
	500	555	555	605
21 year old female	266	304	279	287
	435	430	464	478
30 year old female	177	177	177	213
	320	325	325	368
40 year old male	154	162	162	189
	230	230	230	295

The table shows the motor insurance premiums for various categories of policyholders, with 0,3,6 or 9 points on their driving licenses (note that 3 points are incurred for each speeding conviction). For each category of policyholder, the top row gives the premiums to be paid for third party fire and theft only policies, and the bottom row gives the premiums to be paid for comprehensive policies.

(i) Carry out initial plots to illustrate how the "response variable" premiums paid depends on the four factors Age, Sex, Type of Policy and Number of Points. [3]

(ii) Fit an additive model for this dependence and summarise it. Test the hypothesis that premiums are not affected by the first two speeding convictions. [8]

(iii) What happens if you now include two-factor interactions? How might you improve your model by transforming the response? Summarise how premiums depend on the four factors. [9]

2 The data below show the numbers (N01, N02, N03) of MRSA bacteraemia reports and the corresponding rates (rate 01, rate 02, rate 03) per 1000 bed-days for the time periods April 2001 to March 2002, April 2002 to March 2003 and April 2003 to March 2004 for 45 National Health Service Trusts. (Department of Health Mandatory Bacteraemia Surveillence Scheme). The data have been slightly edited for examination purposes.

	N01	rate01	N02	rate02	N03	rate03
Addenbrookes	110	0.27	127	0.32	126	0.38
Ashford&StPeter	60	0.29	65	0.32	44	0.23
Barts*	62	0.19	74	0.21	62	0.17
Brighton&SussexU	86	0.23	74	0.18	107	0.30
Bucks	39	0.13	43	0.14	47	0.16
CManchester	39	0.11	38	0.12	59	0.17
Chelsea&Westm*	36	0.27	32	0.22	38	0.22
EastKent	99	0.20	85	0.18	70	0.15
Guy's&StThomas*	114	0.32	154	0.43	166	0.45
Hammersmith*	89	0.28	115	0.35	125	0.37
Hull&EYorks	106	0.26	75	0.18	102	0.24
King'sColl*	92	0.31	108	0.37	107	0.35
Lancs	76	0.22	58	0.16	56	0.16
Leeds	196	0.22	165	0.19	204	0.24
LiverpoolWomens	0	0.00	4	0.05	4	0.07
Maidstone&TWells	56	0.21	50	0.19	61	0.24
Medway	NA	NA	29	0.13	48	0.22
NewcastleuTyne	88	0.17	71	0.14	93	0.18
NBristol	144	0.32	114	0.20	88	0.15
NHampshire	29	0.21	13	0.10	20	0.15
NStaffs	83	0.30	87	0.22	135	0.35
NWLondon*	59	0.23	44	0.16	55	0.20
NottinghamC	73	0.25	85	0.29	51	0.17
OxfordRadcliffe	92	0.23	114	0.29	127	0.40
Plymouth	99	0.32	81	0.26	98	0.31
Portsmouth	97	0.32	105	0.33	105	0.32
QuVictoria	6	0.22	5	0.19	3	0.12
QuNottingham	71	0.23	58	0.19	77	0.25
RBerks&Battle	33	0.14	42	0.17	38	0.15
RFreeHampstead*	122	0.41	101	0.39	98	0.34
RSurrey	13	0.10	35	0.23	28	0.18
SalfordR	55	0.24	66	0.28	71	0.25
Sheffield	67	0.11	91	0.15	103	0.16
SManchester	30	0.09	31	0.10	39	0.15
STees	120	0.30	96	0.23	69	0.20
SouthamptonU	45	0.12	53	0.13	62	0.21
SDerbyshire	45	0.14	26	0.09	49	0.15
StGeorges*	115	0.38	75	0.24	93	0.31
StMarys*	64	0.34	72	0.35	59	0.27
RWestSussex	33	0.23	22	0.14	22	0.16
UtdBristol	81	0.29	107	0.37	86	0.28
UCollLondon*	94	0.33	84	0.33	85	0.32
UnBirmingham	189	0.66	169	0.49	123	0.35
UnCoventry&War	74	0.18	82	0.21	79	0.20
UnLeicester	163	0.26	144	0.20	132	0.20

(i) Summarise the data using appropriate plots and tables. [3]

(ii) Use nonparametric methods to test whether or not the MRSA rates have changed between 2001-2 and 2002-3, and between 2002-3 and 2003-4. [3]

(iii) London Hospital Trusts are indicated by an asterisk. Use nonparametric methods to investigate whether there is a difference in rates between London and non-London Hospital Trusts. [6]

(iv) Using Poisson models for the number of reports, investigate the dependence of the rates of MRSA on the year and on whether or not the Trust is in London. Comment briefly on the fit of such models. [8]

3 The data below are taken from the British Medical Journal (March 2005) and show the numbers of cases and deaths over a three year period for 24 surgeons carrying out a particular type of operation in four hospitals, with low- and high-risk patients.

		low risk patients		high risk patients	
Surgeon	Hospital	cases	deaths	cases	deaths
1	1	349	1	76	4
2	2	223	2	35	1
3	2	248	2	42	3
4	2	347	3	53	6
5	3	415	5	112	8
6	3	469	4	98	4
7	1	379	1	69	1
8	3	252	6	56	2
9	3	230	3	63	8
10	4	311	5	51	2
11	4	349	1	64	1
12	2	247	1	19	0
13	2	191	1	48	2
14	4	275	3	53	3
15	3	412	4	76	4
16	1	419	5	84	6
17	4	286	5	51	4
18	3	149	2	48	5
19	4	375	7	63	7
20	3	406	3	107	5
21	3	290	5	81	4
22	1	229	1	51	0
23	2	330	4	56	2
24	2	323	2	65	1

Carry out preliminary summaries and plots for this data set. [4]

By fitting a suitable model and interpreting the output, investigate the mortality rates for these surgeons, taking account of this risk status of the patients. [9]

Investigate whether any observed differences are adequately explained as differences between hospitals. [7]

4 Shown below is a subset of dataset on 500 patients from a GP practice where the body mass index (bmi), age and sex of the patients were recorded.

id	sex	age	bmi
1	1	18	23.39
2	0	47	21.40
3	0	30	30.18
4	0	60	18.58
⋮			
499	0	40	24.29
500	1	25	24.70

id = Patient's anonymous identifier

sex = Sex of patient (0 corresponds to female; 1 to male)

age = Age of patient (in years)

bmi = Body mass index (kg/m^2)

The researchers who collected the data are interested in determining how bmi depends on age and sex, and whether patients, after taking into account their age and sex, can be clustered into two different groups based on their body mass index. They try `lm(bmi~ age + sex)`, and then realize that they do not know how to proceed to the clustering. They approach you with their data.

(a) Plot a histogram of the residuals from the regression, and fit a suitable parametric mixture distribution. You need to provide all relevant plots and details of the mixture model to justify the number of groups you have decided upon. [15]

(b) The researchers also wish to know how they would go about classifying a female patient aged 20 with a body mass index of 23 into one of the groups that you have derived. Use the parametric mixture model that you have derived in (a) to demonstrate how best to allocate this female patient to a group. You need to justify the allocation rule that you use. [5]

END OF PAPER