UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

9am Monday 13 June to 1pm Thursday 16 June 2005

APPLIED STATISTICS

*Attempt **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

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 Monday 13 June, and must hand in their scripts to the Chairman of Examiners by 1.00pm on Thursday 16 June.

The data sets will be emailed to candidates on Monday 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.

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

1 The Times, September 30, 2004, published the data for which the first few lines are given below, under the headline "14 per cent of students drop out of university". The UK universities are listed by Drop-out rate: this is defined as "the percentage of students entering in 2001/02 who are not projected to complete their courses". The "benchmark dropout" represents the expected dropout figure of that institution. This is the 2nd column, Bdropout, in the data matrix. The 4 remaining columns correspond respectively to

(i) the percentage of the State-educated (as opposed to privately educated) for that institution, taken from the 2002/03 university entrants,

(ii) the corresponding Benchmark figure

(iii) the percentage of "poorer" students (i.e. from the lowest 4 social groups of the 7-group socio-economic classification); and

(iv) the corresponding Benchmark figure.

These data were provided by the Higher Education Statistics Agency.

Please answer the following questions.

(a) Give a brief general summary of the data, using appropriate graphs and a paragraph of text, noting any special features.

(b) Compute the 6×6 correlation matrix for the data, and find the correlation of Dropout with State, conditional on the variable Poorer. Interpret your answer.

(c) Define Y as 1 if Dropout $<$ Bdropout, Y as 0 otherwise. How is the binary variable Y dependent on the last 4 columns of the table? Give your best model, with reasons for your answer.

	Dropout	Bdropout	State	Bstate	Poorer	Bpoorer
Cambridge	1.3	3.5	57.6	76.8	11.3	17.3
Oxford	2.1	3.0	55.4	77.2	11.0	17.2
Durham	2.3	6.4	68.3	80.1	15.1	20.4
Nottingham	2.5	5.6	72.8	79.9	16.9	21.0
School_of_Pharmacy,London	3.1	6.9	83.0	88.3	37.0	29.4
London_School_of_Economics	3.2	5.0	66.1	79.3	18.0	19.8
Bath	3.3	6.2	79.7	83.8	16.8	23.9
Bristol	3.4	4.5	63.8	78.9	13.7	20.0
Imperial_College,London	3.6	4.5	62.7	78.2	17.9	19.7
RoyalWelshColl_of_Music&Drama	3.6	11.9	88.5	91.3	23.4	26.9
RoyalCollegeofMusic	3.7	15.3	45.5	89.7	NA	NA
Keele	4.2	11.3	91.2	87.1	25.2	27.9
StranmillisUniv_College	4.4	10.8	99.6	92.7	40.1	33.8
Royal_Holloway&Bedford	4.6	8.7	77.8	83.7	23.5	24.1
StGeorge'sHospMed_School	4.7	6.5	68.8	75.7	31.0	22.5

2 Under the headline "Police spend half their time at a desk: crime detection rates are falling as paperwork burden cuts hours spent on frontline duties", the following table of data was given in The Times, on September 23, 2004. This covers the 42 police forces of England and Wales. Here the column headings are respectively Police force, Burglaries per 1,000 residents, Robberies per 1,000 residents, Vehicle crime per 1,000 residents, % of offences brought to justice, % of police on frontline duty, Days lost per year per officer, Call handling, Road policing, Overall score.

(i) Summarise the data with appropriate graphs, tables, and a paragraph of text.

(ii) How does the "overall score" depend on the first 8 variables? Illustrate the use of the following functions in your solution

stepAIC( ) and boxcox( )

(iii) Now display the cross-tabulation of the two factors "Call handling" and "Road policing". Are those two factors independent?

Dyfed-Powys	4.6	0.1	4.7	44.5	65	10.9	fair	fair	794
Northumbria	17.1	0.9	13.1	27.4	59	8.6	fair	good	719
Suffolk	8.4	0.4	8.6	25.3	64	7.9	poor	good	677
Cumbria	8.5	0.3	8.4	29.3	65	9.9	good	fair	669
Hampshire	9.4	0.6	11.6	19.4	64	8.8	good	good	667
Durham	11.4	0.5	12.0	26.3	59	9.8	fair	good	666
Devon& Cornwall	8.8	0.4	10.4	22.1	65	8.2	fair	fair	654
Wiltshire	10.1	0.5	8.8	28.0	62	12.4	good	good	654
Lancashire	14.9	0.9	11.7	21.3	63	10.1	good	good	640
Kent	11.4	0.7	12.1	23.7	67	11.8	good	good	628
Surrey	8.8	0.6	9.4	17.9	65	9.9	good	good	628
South_Wales	14.7	0.6	22.5	23.3	61	10.2	fair	fair	626
West_Midlands	25.5	4.0	22.2	23.9	62	8.1	fair	excellent	624
Leicestershire	17.2	1.6	15.0	22.8	62	8.4	good	fair	623
Merseyside	22.5	1.7	19.4	20.5	57	11.4	fair	fair	614
Essex	10.3	0.9	13.0	16.9	59	8.2	fair	good	605
Hertfordshire	13.1	0.9	15.7	19.7	60	10.4	good	good	605
West_Mercia	12.2	0.6	10.4	21.5	65	10.4	good	good	604
Dorset	10.1	0.6	12.4	17.8	68	10.7	fair	good	601
Warwickshire	14.9	0.9	14.2	20.9	66	8.0	fair	good	601
South_Yorkshire	25.2	1.1	23.0	18.7	62	8.3	poor	fair	598
Bedfordshire	21.3	1.7	17.8	17.9	70	8.5	good	good	597
Lincolnshire	12.4	0.5	9.9	22.6	60	9.2	fair	good	597
Norfolk	8.5	0.6	10.5	20.7	60	9.2	fair	good	597
Cheshire	13.9	0.7	13.0	18.5	65	9.3	good	good	596
Gwent	14.1	0.5	16.1	30.1	54	13.8	fair	good	596
Staffordshire	14.4	0.9	13.2	23.0	57	10.8	good	excellent	596
Humberside	28.3	2.2	24.6	14.1	64	6.1	poor	poor	588
Metropolitan	21.2	5.5	21.6	12.1	66	8.4	fair	good	588
West_Yorkshire	34.4	1.8	26.7	14.9	68	8.3	fair	poor	588
Cambridgeshire	14.4	1.1	14.9	16.3	66	10.3	fair	good	587
Thames_Valley	16.9	1.1	16.9	17.1	65	8.8	fair	excellent	587
Gloucestershire	13.9	1.0	14.8	22.3	64	9.8	good	good	586
Sussex	12.0	1.0	12.5	18.1	62	11.3	poor	good	583
Derbyshire	19.0	1.1	14.9	18.0	63	9.7	good	good	578
Cleveland	28.7	2.5	23.1	20.7	56	9.0	poor	fair	571
North_Wales	8.9	0.3	11.4	22.5	58	10.5	fair	good	568
North_Yorkshire	15.5	0.5	12.6	18.7	65	11.1	fair	fair	567
Greater_Manchester	36.1	3.7	23.6	16.9	66	10.2	fair	good	555
Northamptonshire	20.1	1.9	19.9	20.7	67	12.2	good	good	549
Avon& Somerset	17.4	2.0	19.5	19.0	62	11.0	good	excellent	542
Nottinghamshire	37.2	2.5	27.7	15.5	60	11.6	fair	fair	539

3 The Times, September 24, 2004, under the headline 'Still waiting for the trains' says 'A fractional improvement in rail punctuality leaves travellers time to ponder results of investment'. The Table shows 'How they did: percentage of trains arriving on time (April to June)'. There are 3 types of train services given in the table below: the first 6 rows correspond to "Long-distance", the next 9 rows correspond to "London&SE", and the final 8 rows correspond to "Regional". The first, second columns of numbers correspond to the percentage of trains arriving on time in the quarter April-June in 2004, 2003 respectively.

	perc04	perc03
FirstGreatWestern	82.1	75.2
GNER	78.4	78.1
MidlandMainline	85.5	68.8
ONE(InterCity)	80.9	81.7
VirginCrossCountry	80.2	70.1
VirginWestCoast	76.3	77.6
c2c	93.1	86.4
ChilternRailways	93.3	91.4
FirstGWlink	84.6	79.3
Silverlink	83.0	88.1
SETrains	86.5	84.2
SWTrains	78.2	78.0
Southern	81.8	84.6
Thameslink	79.4	78.2
WAGN	89.9	86.3
ArrivaTrainsNorthern	88.5	87.0
ArrivaTrainsWales	82.5	84.8
CentralTrains	78.0	76.2
FirstNorthWestern	84.2	85.9
GatwickExpress	82.2	87.9
MerseyRail	95.0	94.8
ScotRail	86.1	87.2
WessexTrains	86.9	85.1

(i) How does the percentage of trains arriving on time depend on the *year* (2003 or 2004) and/or the type of train service (Long-distance, London&SE or Regional)?

(ii) Using appropriate non-parametric methods, test, for all services

(a) whether perc04 is *associated* with perc03,

and

(b) whether perc04 tends to be *bigger* than perc03.

What is the answer to (b) for the London&SE trains only?

4 Shown below is a subset of a dataset from a randomised controlled trial on adult patients (over forty years old) suffering with rheumatoid arthritis. This trial was set up to investigate the effectiveness of Infliximab plus Methotrexate compared with Methotrexate alone in treating rheumatoid arthritis patients. Additionally for a randomly chosen subset of 200 patients (i.e. the data shown below), cost data was also collected. The health economic outcome for the trial was quality adjusted life years (QALY), which is a measure of the value of health that combines length and quality of life into a single index number. Additional information on the age of patients was recorded. It is expected that Infliximab plus Methotrexate will give patients a better quality of life over the trial period, but will be more expensive that receiving Methotrexate alone.

A subset of the data is shown below. The codes for the headers are also presented.

yqaly	ycost	trt	age
10.48	8490.31	1	2
13.8	53272.73	1	2
8.71	21855.2	1	2
14.99	63254.82	1	2
7.32	5606.07	0	2
11.85	36788.09	1	1
9.38	40642.32	1	2
6.58	6857.06	0	1
.			
.			
.			
8.14	41864.31	1	1
8.08	37969.56	1	2
8.51	26933.34	1	1
7.9	3932.14	0	2
7.3	8346.89	0	0
7.35	5953.62	0	2
10.94	19653.58	1	0
11.12	25993.32	0	0
9.07	10752.49	0	0

yqaly = Quality adjusted life years
ycost = Cost of treating patient in pounds
trt = The drug treatment received (0 corresponds to Methotrexate alone; 1 to Infliximab and Methotrexate)
age = Age group of patient (0 corresponds to 40-55 years; 1 corresponds to 55-65 years; and 2 corresponds to 65 years and over)

(a) Produce appropriate plots and summaries of the data by treatment group. Determine whether there are treatment effects on quality adjusted life years and costs.

(b) Given that National Institute for Clinical Excellence (NICE) will reimburse, under the NHS, up to a maximum of £30,000 "per quality adjusted life year gained" for patients with rheumatoid arthritis, determine whether the combination of Infliximab and Methotrexate is cost effective (i.e. good value for money). You should include in your report appropriate graphs and measures of uncertainty on the estimates found.

(c) What are the effects of age and treatment on cost, and on quality adjusted life years? Appropriate reasons, graphs, estimates etc. are required.

END OF PAPER