UNIVERSITY OF CAMBRIDGE

M. Phil. in STATISTICAL SCIENCE

9:00 am Tuesday, 9 June to 1:00 pm Friday, 12 June, 2009.

APPLIED STATISTICS

*Attempt at most **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 9:00 am on 9th June, and must hand in their scripts to the Chairmain of Examiners by 1:00 pm on 12th June. The data sets will be emailed to candidates on 9th June. (The Statistical Laboratory Computer Officer and an Examiner will normally be available for consultation if required between 9:00 am and 4:30 pm on 9th, 10th and 11th June and between 9:00 am and 1:00 pm on 12th June.)

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. Candidates are advised to state models algebraically and to discuss formally the details of their statistical analyses.

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 In the nineteenth century, data were collected from all the Departments in France (source: Guerry (1833)) and the first few lines of a subset of the data for 85 Departments are shown below. The variables are

`dept`: the identity number for the Department (the Departments are subregions of France, not all Departments are included in this dataset);
`Region`: the regions of France (`N`, `E`, `S`, `W` and `C` are North, East, South, West and Central, respectively);
`CrimesPers`: the Department poplulation divided by the number of crimes against the person committed in the Department during the period 1825–1830;
`CrimesProp`: the Department population divided by the number of crimes against property committed in the Department during the period 1825–1830;
`Literacy`: the percentage of military conscripts who could read and write;
`Donations`: the total amount of donations to the poor;
`Wealth`: A ranked index based on taxes on personal and moveable property per inhabitant (`1` corresponds to the maximum).

dept	Region	CrimePers	CrimeProp	Literacy	Donations	Wealth
1	E	28870	15890	37	5098	73
2	N	26226	5521	51	8901	22
3	C	26747	7925	13	10973	61
4	E	12935	7289	46	2733	76
5	E	17488	8174	69	6962	83
7	S	9474	10263	27	3188	84

(a) Summarise the data using appropriate tables, plots and summaries.

(b) Define a new variable `y` to be the ratio `CrimePers/CrimeProp`. What is `y` in terms of the number of crimes against the person and the number of crimes against property? By fitting and comparing suitable models, investigate how `y` depends on `Region`, `Literacy`, `Donations` and `Wealth`, illustrating the use of `boxcox` in your answer. Summarise your results in words.

2 In vitro fertilisation (IVF) data for 1992–2005 show the number of IVF treatment cycles, together with the number of singleton births, twin births and triplet and higher order births, for each year (source: HFEA). The data are shown in the table below. For example, out of 18201 cycles of IVF treatment in year 1, there were 2373 (= 1712+591+70) pregnancies leading to live births, of which 1712 were singleton births, 591 were twin births (each resulting in a pair of babies), and 70 resulted in three or more babies.

Year	Number of treatment cycles	Singleton births	Twin births	Triplet and higher order births
1	18201	1712	591	70
2	21239	2244	738	110
3	23517	2391	837	123
4	25414	2589	915	106
5	27203	3015	1041	123
6	25033	2781	888	113
7	23551	2812	978	113
8	22737	2945	1013	74
9	22720	3083	1002	81
10	22342	3116	1007	53
11	22477	3284	1096	33
12	21884	3371	1043	25
13	23250	3460	1015	15
14	23794	3626	1132	15

(a) Consider the probability that a treatment cycle gives rise to a singleton birth. Use a χ^2 test to determine whether or not the data are consistent with this probability being the same for all fourteen years.

(b) Use binomial models to investigate how the probability that a treatment cycle results in a singleton birth depends on the year, illustrating your answer with appropriate plots and providing estimates and confidence intervals for the relevant parameters. According to your preferred model, explain how the odds in favour of a singleton birth change from one year to the next. Comment on the fit of your model.

(c) Consider the probability that a birth results in a multiple birth, ie that a birth results in two or more babies. Using appropriate models, investigate how this probability depends on year.

3 Twenty providers of bus services in rural areas of the USA were questioned about their passenger numbers and about other quantities that may be useful in predicting passenger numbers. A few lines of the resulting data are shown below (Journal of Transportation Engineering, edited for examination purposes):

Passengers	Bmiles	Older	Fare	Pop	LIH	Poverty	MedRent
68	160	1166	3.25	6007	234	329	313
80	137	2285	0.75	12000	827	2040	435
85	130	1960	0.25	7637	827	1828	299
110	22	450	0.00	3515	139	217	271

The variables in the dataset are:
`Passengers`: bus passengers (mean number per day);
`Bmiles`: the total daily miles travelled by the provider's buses;
`Older`: the number of people over 55 in the area served by the bus provider;
`Fare`: the fare for a one-way trip in dollars;
`Pop`: the population in the area served by the bus provider;
`LIH`: the number of households classified as low income in the area served by the bus provider;
`Poverty`: the number of households classified as below the poverty line in the area served by the bus provider;
`MedRent`: the median rent in the area served by the bus provider.

(a) Carry out exploratory plots and summaries of these data, commenting briefly on what you find.

(b) Using Poisson models, investigate how `Passengers` depends on the other variables.

(c) In fact, the first eight of the providers in the dataset were providing a fixed-route service and the remaining twelve were providing an on-demand service. By including in your models a suitable two-level factor identifying the type of service provider, investigate how `Passengers` depends on the other variables, and on whether this dependence is different for the two types of provider.

4 In the Isle of Lagend, village communities believe that violence is on the increase over the last five years. In order to determine whether there is any evidence for this belief, Tlacatecuhtli (Leader of Lagend) instructs the Heads of the villages to collate the data on Tommi-puukko (a form of knife) incidents over the five-year period from 2004 to 2008. Additionally, Tlacatecuhtli asks for information on the village's population size, the proportion of the village members who are male aged 14 to 24 years, the village's size (in terms of land area covered) and the location of the village (whether near the coast or further inland).

Shown below is a subset of the data collected by the Heads of the 32 villages.

Village	Time	TPattacks	PopnSize	Inland	VillSize	PropMale
1	0	21	315	0	2	27.4
1	1	23	322	0	2	27.4
1	2	25	308	0	1	27.4
1	3	30	320	0	1	27.4
1	4	32	326	0	1	27.4
.						
.						
.						
32	0	32	279	1	1	40.0
32	1	29	264	1	1	40.0
32	2	25	272	1	1	40.0
32	3	38	273	1	2	40.0
32	4	52	265	1	2	40.0

`Village`: The Study Identification Numbers given to villages

`Time`: The year for which the Tommi-puukko episode data applies ($0, 1, \ldots, 4$ corresponds to $2004, 2005, \ldots, 2008$ respectively)

`TPattacks`: Tommi-puukko attacks

`PopnSize`: Population size of a village

`Inland`: Location of village (0 corresponds to Coastal; 1 corresponds to Inland)

`VillSize`: Size of Village (1 corresponds to Small; 2 to Medium; 3 to Large)

`PropMale`: Proportion of a village which is young males (between 14 and 24 years old) taken at the 2006 census.

In discussing the information collated, the Village Heads suspect that each village may have a different propensity for Tommi-puukko incidents and that there may be heterogeneity in the annual rate of change in attacks across villages. Tlacatecuhtli suspects that most of these incidents involve young males and occur in villages that are further inland. However there is uncertainty on what role the size of the village, and even some form of interaction of it with the population size of the village, plays in influencing attacks.

In order to analyse the data collected, Tlacatecuhtli decides to contract out the statistical work to you. The specific aims of this project, as set out by Tlacatecuhtli and the Village Heads, are

(a) To describe graphically how the rates of Tommi-puukko incidents change over the five-year period for each village.

(b) To determine, based on the data across all the villages, whether marginally there is an increasing or decreasing trend in the rates of Tommi-puukko incidents over the five years, and if so, to estimate and interpret the size of the trend.

(c) To investigate the variation in the rates of Tommi-puukko attacks in the year 2008 and the heterogeneity in the annual rates of change in attacks across villages, whilst ignoring any explanatory variables that could potentially explain them.

(d) To investigate the associations between the explanatory variables collected and the rate of Tommi-puukko incidents, by fitting appropriate longitudinal data models.

(Note that Tlacatecuhtli has suggested categorizing the PropMale variable into three categories: Low, Medium, and High, based on cut-points of 20% and 30%; and dichotomizing the PopnSize variable, *if treated as a covariate*, into population sizes above 250 and not above 250, for purposes of simplicity.

By the appropriate modelling of the data and use of plots, analyse the data in order to address the objectives of Tlacatecuhtli and the Village Heads. Inferences on the effects of the explanatory variables on the rate and their interpretations must be made from the final multivariate model that you have decided "best" fit the data. You need to justify statistically your choice of "best" model. Is there any evidence to substantiate Tlacatecuhtli's suspicion that young males are involved in most of these Tommi-puukko incidents?

END OF PAPER