UNIVERSITY OF CAMBRIDGE

M. Phil. in STATISTICAL SCIENCE

9:00 am Tuesday, 8 June to 1:00 pm Friday, 11 June, 2010.

APPLIED STATISTICS

*Attempt no more than **THREE** questions.*

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

The questions carry equal weight.

This is an 'Open Book' examination. Candidates will receive this paper at 9:00 am on 8th June, and must hand in their scripts to the Chairmain of Examiners by 1:00 pm on 11th June. The data sets will be emailed to candidates on 8th 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 8th, 9th, and 10th June and between 9:00 am and 1:00 pm on 11th 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

The Happy Planet Index is an index calculated for each of 143 countries in 2005, based on the life expectancy, the life satisfaction and the ecological footprint for each country. The first few lines of the data are given below.

Country	Subregion	LifeExp	LifeSat	Footprint	HPI
Albania	7b	76.2	5.5	2.2	47.9
Algeria	3a	71.7	5.6	1.7	51.2
Angola	4a	41.7	4.3	0.9	26.8
Argentina	1b	74.8	7.1	2.5	59.0
Armenia	7a	71.7	5.0	1.4	48.3

The variables are:

`Country`: The 143 countries in alphabetical order.

`Subregion`: Each country belongs to a subregion, which is itself in a region. These are defined as follows.

Region	Subregion	
1	1a	Central America, Mexico and Carribean
	1b	South America
	1c	Uruguay
2	2a	Australia and New Zealand
	2b	North America
	2c	Western Europe
	2d	Nordic Europe
	2e	Southern Europe
3	3a	North Africa
	3b	Middle East, South West Asia
4	4a	Southern and Central Africa
	4b	East Africa
	4c	West Africa
5	5a	South Asia
6	6a	China
	6b	Wealthy East Asia
	6c	South East Asia
7	7a	Central Asia and Caucasus
	7b	Central and Eastern Europe
	7c	Russia, Ukraine and Belarus

`LifeExp`: Life expectancy at birth in years in 2005.

`LifeSat`: A measure of life satisfaction between 0 and 10, where 0 is dissatisfied and 10 is satisfied.

`Footprint`: The per capita footprint (in global hectares). This is a measure of the amount of land required for each person's resource requirements plus the amount of vegetated land needed to absorb each person's CO_2 emissions (including emissions for the products the person consumes). A person using up to 2.1 global hectares is regarded as using their fair share of the world's resources.

`HPI`: The Happy Planet Index.

(a) Summarise the data using summary statistics, plots and tables as appropriate.

(b) Using linear models, investigate how the Happy Planet Index is related to `log(LifeExp)`, `log(LifeSat)` and `log(Footprint)`, and include consideration of possible transformation of the response variable. Test whether the two variables `LifeExp` and `LifeSat` may be replaced by the variable *happy life years*, obtained as `LifeExp` $*$ `LifeSat`/10.

(c) Investigate how life satisfaction depends on life expectancy, the ecological footprint, and on the region.

[*Hint: Use the function* `as.numeric` *or otherwise to construct the regions.*]

2

The data below (edited for examination purposes) concern airline delays for charter flights to nine UK airports in the summer seasons (April to October) of 2006 and 2007.

airline	nationality	nflights06	nlate06	nflights07	nlate07
1	Spain	197	15	197	24
2	UK	19111	1652	18026	1796
3	UK	9759	895	10022	1042
4	Spain	281	26	297	23
5	Malta	673	82	643	64
6	UK	10900	1017	9665	986
7	UK	15538	1367	15418	1634
8	UK	7088	652	6453	683
9	UK	2257	231	1423	165
10	Spain	540	47	153	21
11	UK	15248	1357	15172	1503
12	Spain	1471	193	164	24
13	Cyprus	928	169	882	135
14	UK	651	71	482	57
15	Turkey	1291	178	1993	369
16	UK	NA	NA	859	107

The variables are

`airline`: There are 16 airlines, coded `1, ... , 16`.

`nationality`: The nationality of the airline.

`nflights06, nflights07`: The number of charter flights to the nine UK airports in the summers of 2006 and 2007 respectively.

`nlate06, nlate07`: The number of charter flights to the nine UK airports in the summers of 2006 and 2007 respectively that were more than one hour late.

(a) For 2007, use χ^2 tests to test whether there is any difference between the probability of a flight being more than one hour late

  (i) for the airline from Malta compared to all airlines from Spain combined,

  (ii) for all UK airlines combined compared to non-UK airlines combined.

(b) An airline is classed as *busy* if it has more than 5000 charter flights to these nine UK airports in the summer season of 2007, and is *non-busy* otherwise. Using appropriate Poisson modelling for the 2007 data, investigate how the probability of a flight in 2007 being more than one hour late depends on the nationality of the airline and on whether the airline is busy or not.

(c) Using Poisson models, investigate how the probability of a flight being more than one hour late depends on the airline and on the year. Using your preferred model, estimate what the number of late flights for airline 16 in 2006 would have been if that airline had had the same number of flights in 2006 as in 2007.

3

The data below refer to the educational backgrounds of members of the UK Houses of Parliament in 2005 in the three main political parties. For those members for which information was available, the number of those who went to independent schools (ie to fee-paying private schools) is recorded below, in addition to the number who went to university, and the number who went to Oxford or Cambridge Universities.

House	Party	Type	nschooltot	nindep	nunivtot	nuniv	noxb
Comm	Lab	Office	89	22	87	62	20
Comm	Lab	Back	230	36	254	165	37
Comm	Con	Office	82	51	81	67	37
Comm	Con	Back	101	57	111	92	51
Comm	Lib	Office	35	13	35	30	12
Comm	Lib	Back	21	9	27	20	7
Lords	Lab	NA	176	60	194	138	51
Lords	Con	NA	193	153	186	132	92
Lords	Lib	NA	65	35	70	54	29

In this dataset the variables are

`House`: A member of the Houses of Parliament is either a member of the House of Commons (`Comm`) or of the House of Lords (`Lords`).

`Party`: The three main political parties are Labour (`Lab`), Conservative (`Con`) and Liberal Democrat (`Lib`).

`Type`: For members of the House of Commons only, this variable is `Office` if the member held office within his or her political party, and is `Back` if the member is a backbencher (ie did not hold office).

`nschooltot`: The number of members for which data on type of school is available and who were at school in the UK.

`nindep`: The number of members out of those in `nschooltot` who attended an independent school.

`nunivtot`: The number of members for which information is available on whether or not they attended university, and who, if they went to university, were at university in the UK.

`nuniv`: The number of members out of those in `nunivtot` who have been to university.

`noxb`: The number of members out of those in `nunivtot` who attended Oxford or Cambridge University.

Using relevant preliminary plots and summaries, and using appropriate models and statistical tests, investigate the following questions about members of the main three political parties for whom information is available:

(a) How does the type of school of a member depend on the factors `House` and `Party`?

(b) How does whether or not a member went to university depend on the factors `House` and `Party`?

(c) For members of the House of Commons who went to university, does whether or not a member went to Oxford or Cambridge depend on being an officeholder or backbencher when differences due to political party have been taken into account?

4

Teenage pregnancy is a serious social problem in the United Kingdom. The teenage conception rates in the UK are amongst the highest in the EU countries. Having children at a young age can damage young women's health and well-being, and severely compromise their education and career prospects. While individual young people can be competent and responsible parents, evidence-based research shows that children born to teenagers are much more likely to experience a range of negative outcomes in later life.

The UK Government recognised that action to reduce the number of teenage pregnancies in the UK was needed in order to address the social, health and economic consequences of a high teenage conception rate. In 1999 the UK Government launched a Teenage Pregnancy Strategy which required all Local Authorities in England and Wales to reduce their conception rate in Under 18s by 50% from their 1998 figure by 2010.

Shown below is a subset of the data (suitably edited) obtained from the Office for National Statistics on conception rates (per 1000) for the Under 18s for the period 1998 to 2006 by the 374 Local Authorities in England and Wales (excluding City of London and Isles of Scilly).

gorname	laname	year	count	rate
North East	Gateshead	0	199	57.1
North East	Newcastle upon Tyne	0	258	52.8
North East	North Tyneside	0	204	58.4
⋮	⋮	⋮	⋮	⋮
Wales	Monmouthshire	8	62	35.0
Wales	Newport	8	151	50.3
Wales	Cardiff	8	263	58.4

`gorname`: Name of the Geographical Region in England and Wales

`laname`: Name of the Local Authority

`year`: Variable (coded 0 to 8) denoting the years (1998 to 2006 respectively) to which the conception rates apply. Therefore (year + 1998) gives the actual Calendar Year

`count`: Number of Under 18s conceptions

`rate`: Rate of Under 18s conceptions (per 1000 females between ages 15 and 17 inclusive)

By appropriate modelling of the data and use of plots, fit an appropriate longitudinal data model to the conception rates in order to answer (a) to (e) below.

(a) Interpret the fixed effects estimates obtained from your model.

(b) Determine how many and which Local Authorities would be predicted to meet the Government's target of a halving of the 1998 conception rate by 2010.

(c) Determine which additional Local Authorities would be predicted to meet the target if the target is revised to a 30% reduction from the original 50% reduction.

(d) Characterise the variation in Under 18s conception rates across the Local Authorities at baseline (i.e. 1998) and the heterogeneity of the rate of change in conception rates over time.

(e) Determine what the variation in Under 18s conception rates across the Local Authorities would be in 2010.

Would it be statistically defensible to rank the Local Authorities by their predicted Under 18s conception rates for 2010? Give reasons for your answer.

END OF PAPER