Examination Practical Session

As well as your written answers to these questions, your performance will be assessed by the log files and do files that you produce. Please ensure that your do file is edited so that it will run straight through without stopping. The do file should write a log-file to P:/stats_exam.

You are strongly recommend to start a command log with the command

```
cmdlog using P:/stats_exam/exam.do
```

This will ensure that all of the commands you enter are kept. You can then edit exam.do to produce your final result (you will need to enter the command

```
cmdlog close
```

before trying to edit exam.do).

To ensure that a log of your results is kept, the first two commands in the do-file should be

```
capture log close
log using P:/stats_exam/exam.log, text replace
```

These commands should be followed by a comment containing your name, so that I know which log-file was produced by which individual. Comments are lines that begin with an asterisk: stata does not attempt to process commands from these lines, but puts them directly into the log file unaltered. So, to add your name, type

```
* Your Name
```

Answer all questions in your log file. Again, this can be done by adding comments to the do-file, For example, to put the answer “Yes” to question 2.4, you would type

```
* 2.4 Yes
```

All of the datasets required for this exam can be found at

```
http://personalpages.manchester.ac.uk/staff/mark.lunt/data/exam_2008
```

You can either work with the data from there, reading it into stata with commands of the form

```
use http://...
```

or copy the data to P:/stats_exam and read the data in from there.

1 Data Manipulation

For this section, the data you need is in the dataset `bmd.dta`. This contains some data taken from the EVOS study of bone density and vertebral fracture in a number of European centres.

1.1 The variable `gender` contains 0 for males and 1 for females. Use `label define` to produce a suitable label for this variable, and apply it to the variable

1.2 Create a new variable called `logspi` which contains the log of the spinal BMD (in `cal_spi`).

1.3 Use `label variable` to apply a suitable label to this variable.

1.4 Create a variable called `bmi` which contains the Body Mass Index, defined as

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in metres})^2}$$

1.5 Give a suitable label to the variable `bmi`.

1.6 Create a variable called `age2` containing the square of the age, and label it appropriately.

1.7 Save this dataset to the P:/stats_exam using the name “new_exam”

2 Descriptive Statistics

This data for this part in the file `iq.dta`. This file contains data concerning 3 measures of IQ, and the brain volume (in pixels) measured by MRI. The subjects height, weight and gender are also recorded.

2.1 How many observations are there in the dataset ?

2.2 How many observations are on male subjects ?

2.3 How many subjects have missing data for their height ?

2.4 What is the mean Full Scale IQ score ?

2.5 What are the median Verbal IQ scores in men and in women ?

2.6 Draw a histogram of Full Scale IQ score: is it normally distributed ?

3 Linear Regression

3.1 Create a linear regression model in which brain size is predicted from Full Scale IQ. Is there a significant association ?

3.2 What proportion of the variance in brain size is explained by Full Scale IQ ?

3.3 What difference in brain size would be expected between two subjects whose Full Scale IQ differed by 10 points ?

3.4 Add height to the regression equation. Are both height and Full Scale IQ significant predictors of brain size ?

3.5 What proportion of the variance in brain size is explained by Full Scale IQ and height ?

3.6 What would be the expected brain size of a person 70 inches tall and with a Full Scale IQ of 115 ?

3.7 Can you give a 95% confidence interval for your expected brain size ? (Hint: use `lincom`).

3.8 Add gender to your model. Is there a significant difference in brain size between men and women, after adjusting for height and Full Scale IQ ?

3.9 Add an interaction term to your model to test whether the change in brain size with height differs between men and women. What would you conclude ?

4 Regression Diagnostics

This section uses the data from the file `smoking.dta`. This contains data concerning the relative mortality from lung cancer in 25 different occupational groups. Mortality is given as a standardised mortality ratio, where 100 represents mortality the same as in a standard age-matched population, values over 100 represent increased mortality and values less than 100 represent decreased mortality. The variable smoking contains a standardised smoking ratio, measuring how much men in that particular group smoke compared to the standard population.

4.1 Fit a linear regression model to predict mortality from smoking. Is the association statistically significant ?

4.2 Produce a plot of the residuals against the predicted values from the regression model. Is there any evidence of non-constant variance ?

4.3 Perform a formal test for constancy of variance. Does this confirm your previous answer ?

4.4 Produce a scatter plot of mortality against smoking. Is there any evidence of non-linearity ?

4.5 Perform a formal test for non-linearity. Does this confirm your previous answer ?

4.6 From the scatter plot, are there any outliers in the data ?

4.7 Are the residuals from the regression normally distributed ?

5 Logistic Regression

This section uses data from a study of survival in an Intensive Care Unit. The outcome variable is `died` which measure whether the patient died before they could be discharged from hospital. A large number of potential predictors were measured when the subject was admitted, to determine how well it is possible to identify subjects with a poor probability of surviving. Only a small selection of the subjects and variables are included in this dataset.

5.1 What proportion of subjects in this dataset died ?

5.2 What proportion of emergency admissions died ?

5.3 Fit a logistic regression model to predict probability of dying from gender. Does the probability of survival differ between men and women ?

5.4 Give the odds ratio for dying, for men compared to women, along with its 95% confidence interval. **Note:** check the coding of sex: it differs from that for gender in the brain size dataset.

5.5 Is the risk of dying associated with associated with age ?

5.6 What is the odds ratio (with its 95% confidence interval) for a 1 year increase in age ?