Examination Practical Session 2011

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/exam.log`".

You are strongly recommended 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
```

The exam requires the use of the datasets
`http://personalpages.manchester.ac.uk/staff/mark.lunt/data/bmd.dta`
`http://personalpages.manchester.ac.uk/staff/mark.lunt/data/iq.dta`
`http://personalpages.manchester.ac.uk/staff/mark.lunt/data/icu.dta`
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: either will work when I run your do-file.

1 Data Manipulation

This section uses the dataset `bmd`, which contains data on spine bone mineral density (BMD) on 2,416 subject. The variables in the dataset are

centre Centre number

id Individual's ID number

cal_spi Spinal BMD in g/cm^2

gender The individual's gender

height Height in metres

weight Weight in kg

age Age in years

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 (2)

1.2 Create a new variable called `logspi` which contains the log of the spinal BMD (which is in `cal_spi`). (1)

1.3 Use `label variable` to apply a suitable label to this variable. (1)

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)

1.5 Give a suitable label to the variable `bmi`. (1)

1.6 Create a new variable called `age_cat` which contains the value 0 for individuals less than 60 years old, 1 for individuals 60 years old or older. (3)

1.7 Save this dataset to `P:/stats_exam` using the name "new_exam" (1)

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 ? (1)

2.2 How many observations are on male subjects ? (1)

2.3 How many subjects have missing data for their height ? (1)

2.4 What is the mean Full Scale IQ score ? (1)

2.5 What are the median Verbal IQ scores in men and in women ? (2)

2.6 Draw a histogram of Full Scale IQ score: is it normally distributed ? (2)

2.7 Draw boxplots of `mri_count` for men and women. Does there appear to be a difference between the genders ? (2)

2.8 Give the median, 25th and 75 percentile of Full Scale IQ score in men and women separately. (2)

3 Linear Regression

In this section, we are concerned with fitting and interpreting linear regression models. Do not worry about the assumptions underlying linear regression: answer the questions as if you had tested the assumptions and found that the data satified them. We will test them for real in the next section.

This section uses the `bmd` dataset used in Section 1.

3.1 Perform a linear regression that will predict spinal BMD (`cal_spi`) from `age`, `gender`, `height` and `weight`. Which of the four variables are statistically significant predictors of BMD ? (2)

3.2 What is the proportion of the total variance of BMD that can be explained by these predictors ? (1)

3.3 What is the mean difference in BMD between men and women after adjusting for age, height and weight ? Give a 95% confidence interval around this value. (2)

3.4 What is the predicted BMD for a man aged 65, with a weight of 70 kg and a height of 1.7m ? (1)

3.5 Add a categorical predictor to your model to test if there are differences between the centres. Which centre number has the lowest BMD on average ? (2)

3.6 Use `testparm` to determine whether the differences between centres are statistically significant. (1)

3.7 What proportion of the variance in BMD is explained by this model ? (1)

3.8 Add an interaction term to test whether the change in BMD with weight differs between men and women. Is there a statistically significant difference ? (2)

4 Regression Diagnostics

4.1 Use the command `qnorm` to produce normal plots of `cal_spi` and `logspi`. Which of them is more nearly normally distributed ? (3)

4.2 Fit a linear regression model with either `cal_spi` or `logspi` as outcome, whichever was closer to normality in the previous question. Use age, height and weight as continuous predictors and gender as a categorical predictor.Produce a plot of the residuals against the predicted values from the regression model. Is there any evidence of non-constant variance ? (5)

4.3 Produce a component-plus-residual plots for `age`. Is there any evidence of non-linearity ? (Hint: this can be difficult to see: I suggest that you add the option "`lowess`" to the command to get an idea of how the mean changes with weight without assumptions about the form of the association). (2)

4.4 Create a variable `age2` containing the square of `age` and add it to the regression. Is it significant, and what does this tell you about the linearity of the association between `age` and BMD. (3)

4.5 Produce a normal plot of the residuals from the regression model including `age2`. Are they normally distributed ? (2)

4.6 Calculate Cook's distance for each observation, and plot it against the predicted values. Are there any outliers in this plot ? (2)

4.7 Repeat the regress after excluding the 2 observation with the largest values of Cook's distance. Does removing these observations change your conclusions as to which variables are significantly associated with BMD ? (2)

5 Logistic Regression

This section uses data from a study of survival in an Intensive Care Unit contained in the dataset `icu.dta`. 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 ? (1)

5.2 What proportion of emergency admissions died ? (1)

5.3 Fit a logistic regression model to predict survival from gender. Does the probability of survival differ between men and women ? (2)

5.4 Give the odds ratio for survival, 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 dataset `iq.dta`. (2)

5.5 Test whether survival is associated with the type of admission. Is there are statistically significant difference between the admission types ? (1)

5.6 Which type of admission is used as the reference category ? (1)

5.7 Is survival significantly associated with age ? (1)

5.8 Give the odds ratio, along with its 95% confidence interval, for the effect of a 1 year increase in age on the risk of dying. (2)

5.9 Perform a Hosmer-Lemeshow test with 10 groups. Does this test suggest that the fit of this model is adequate ? (2)

5.10 If a Hosmer-Lemeshow test suggests that your model is not adequate, what can you do to improve the fit of the model (you do not need to fit a better model, just say what steps you could take). (3)

The number of marks available for each question is given in parentheses after the question. There are a total of 69 marks on the exam: I will multiply the number of marks by 1.45 to give you a percentage.