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_2009
```

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 Descriptive Statistics

For this section, the data you need is in the dataset `bsrbr.dta`. This contains some data taken from the BSRBR register, which is looking at the safety of anti-TNF drugs in RA. The data contains measurements of HAQ score, disease activity score (DAS), age at recruitment, disease duration and gender on subjects receiving either anti-TNF drugs (`treated == 1`) or conventional DMARDS (`treated == 0`).

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

1.3 How many observations are there in the dataset ?

1.4 How many observations are on male subjects ?

1.5 How many subjects have missing data for their HAQ ?

1.6 What is the mean disease duration ?

1.7 What are the median ages in men and in women ?

1.8 Draw a histogram of DAS in untreated subjects: is it normally distributed ?

1.9 Draw boxplots of DAS for treated and untreated subjects. Does there appear to be a difference between the two groups ?

1.10 Give the median, 25th and 75 percentile of HAQ in treated and untreated subjects separately.

2 Linear Regression

For this next section, we are only concerned with the untreated subjects, so removed the treated subjects with the command

```
drop if treated == 1
```

2.1 Create a linear regression model in which DAS is predicted from HAQ. Is the association significant ?

2.2 What proportion of the variance in DAS is explained by this variable ?

2.3 What difference in DAS would you expect to see between two subjects whose HAQs differed by 0.5 ?

2.4 Add age to the regression model. Are both age and HAQ significant predictors of DAS ?

2.5 What would be the expected DAS of a person aged 50 with a HAQ score of 1.5 ?

2.6 Give a confidence interval for the answer to the previous question.

2.7 Add gender to your model. Is there a significant difference in DAS between men and women after adjusting for age and HAQ ?

2.8 Add an interaction term to your model to test whether the change in DAS with HAQ differs between men and women. Is this term significant ?

2.9 What is the change in DAS per unit change in HAQ in men ?

2.10 What is the change in DAS per unit change in HAQ in women ?

3 Regression Diagnostics

In this section we are going to producing regression diagnostics for the model

```
xi:  regress dascore haq age i.gender
```

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

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

3.3 Produce component-plus-residual plots for age and HAQ. Is there any evidence of non-linearity for either of these variables ?

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

3.5 What are the 2 largest values of Cook's distance ?

3.6 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 the DAS ?

3.7 Produce a normal plot of the residuals. Are the residuals from the regression normally distributed ?

4 Logistic Regression

This next section requires data on both treated and untreated subjects, so reload the entire dataset before you proceed.

In this section we are going to be calculating propensity scores, that is the probability of receiving anti-TNF treatment (`treated == 1`) given the values of the other variables at baseline.

4.1 How many subjects are in the treated and untreated groups ?

4.2 What is the mean age in the treated and untreated groups ?

4.3 Fit a logistic regression model with treatment as the outcome variable and age as the predictor. Is there a statistically significant association between age and the probability of receiving anti-TNF treatment ?

4.4 What is the odds ratio for a one year increase in age, with its 95% confidence interval ?

4.5 Now add HAQ, DAS, disease duration and gender to the logistic regression equation. Which of these variables are significantly associated with receipt of anti-TNF treatment ?

4.6 Perform a Hosmer-Lemeshow test with 10 groups. Is the fit of this model adequate ?

4.7 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 could be done).

4.8 Create a new variable, `propensity`, containing the predicted probability of receiving treatment for each subject.

4.9 What is the median and IQR for `propensity` in subjects receiving anit-TNF treatment and in subjects not receiving anti-TNF treatment ?

4.10 Create a new variable, `pwt`, containing 1/`propensity`.

4.11 Replace `pwt` with $1/(1 - \texttt{propensity})$ for subjects with `treated == 0`.