Examination Practical Session 2010

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
```

All of the questions require the use of the dataset

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

All of the questions involve the same dataset, which is information about crash testing of motor vehicles, and the variables are either characteristics of the car which may affect the amount of damage suffered by and occupant. First, there are 4 variables that identify the car by make, model and year of manufacture. Then there are 4 outcome variables, measuring the forces registered on the crash test dummies at difference sites. The last 6 variables are potential predictors of the severity of damage:

location Location of occupant (Driver or Passenger)

protection Type of protection provided (seat belts, airbags etc).

doors Number of doors

weight Weight of vehicle

type Size and type of vehicle

year Year of manufacture

1 Data Manipulation

1.1 The variable `location` takes the value 1 if the results apply to the driver, and 2 if the results apply to the passenger. Create a suitable label for the values of this variable and apply it to the variable. (2)

1.2 Create a new variable called `wt2` which contains the square of the the weight, and label it appropriately. (2)

1.3 Create a new variable called `lh` which contains the natural logarithm of Head Impact Criterion variable (`headic`), and label it appropriately (you need to use the function `ln()`). (2)

1.4 Create a new variable containing the average of the left leg and right leg impact variables, and call it `leg_mean`. (1)

1.5 Create a new variable called `bad_leg` which takes the values 0 if `leg_mean` $<$ 1000, 1 if `leg_mean` $\geq$ 1000 and . if `leg_mean` is missing (1)

1.6 Set the default reference category for the predictor `protection` to be category 4 ("manual belts"). (1)

2 Descriptive Statistics

2.1 How many observations are there in the dataset ? (1)

2.2 How many observations are on drivers ? (1)

2.3 How many subjects have missing data for their Head Impact Criterion ? (1)

2.4 What is the mean Chest Deceleration ? (1)

2.5 What are the median Head Impact Criteria in drivers and in passengers ? (2)

2.6 Draw a histogram of Head Impact Criterion: is it normally distributed ? (2)

2.7 Draw boxplots of Chest Deceleration for drivers and passengers. Does there appear to be a difference between the two groups ? (2)

2.8 Give the median, 25th and 75 percentile of Chest Deceleration in drivers and passengers 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.

3.1 Create a linear regression model in which Head Impact Criterion is predicted as a linear function of weight. Is the association statistically significant ? (2)

3.2 What proportion of the variance in Head Impact Criterion is explained by the weight of the vehicle. (1)

3.3 Give the difference in mean Head Impact Criterion would you expect to see between two vehicles whose weights differed by 100 lb, with a 95% confidence interval. (2)

3.4 Add the variable `location` to the regression equation to test whether there are differences in Head Impact Criterion according to whether the occupant is a driver or a passenger. Which variables are now significant predictors of Head Impact Criterion ? (2)

3.5 Give the expected Head Impact Criterion for the driver of a vehicle weighing 3000 lb, along with its 95% confidence interval. (2)

3.6 Give the expected Head Impact Criterion for the passenger in a vehicle weighing 3000 lb, along with its 95% confidence interval. (2)

3.7 Add an interaction term to the model to test whether the change in Head Impact Criterion with weight differs between drivers and passengers. Is there a statistically significant difference ? (2)

4 Regression Diagnostics

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

4.2 Using whichever variable from the previous question is more nearly normally distributed, fit four separate linear regressions, in which the predictor variables are

1. `weight` (as a continuous variable)
2. `location` (as a categorical variable)
3. `protection` (as a categorical variable)
4. `doors` (as a categorical variable)

Which of these four variables have a statistically significant association with Head Impact Criterion ? (2)

4.3 Fit a single regression model including all of the variables that were significant in the previous question. Produce a plot of the residuals against the predicted values from the regression model. Is there any evidence of non-constant variance ? (5)

4.4 Perform a formal test for constancy of variance. Does the result of this test confirm your previous answer ? (2)

4.5 Produce a component-plus-residual plots for `weight`. 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)

4.6 Add the variable `wt2` that you created in section 1 to the regression. Is it significant, and what does this tell you about the linearity of the association between `weight` and `headic`. (3)

4.7 Produce a new component-plus-residual plot. Does this plot suggest that we have the form of the association between `weight` and `headic` correct now ? (2)

4.8 Produce a normal plot of the residuals from the regression model including `wt2`. Are they normally distributed ? (2)

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

4.10 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 Head Impact Criterion ? (2)

5 Logistic Regression

In this section we will be considering a dichotomous outcome variable, `bad_leg`.

5.1 For how many subjects does `bad_leg` take the value 0, and for how many of them does it take the value 1 ? (2)

5.2 What is the mean weight of vehicle for the for drivers with `bad_leg` == 0 and for drivers with `bad_leg` == 1 ? (2)

5.3 Fit a logistic regression model with `bad_leg` as the outcome and `protection` as a categorical predictor variable. Does the risk of a bad outcome vary significantly between the different types of protection ? (2)

5.4 Which type of protection is being used as the reference category ? (Hint: look at question 1.6 (1)

5.5 What is the odds ratio, and its 95% confidence interval, for `bad_leg` == 1 for a motorized belt, relative to reference category ? (2)

5.6 Add the predictor `weight` to the logistic regression equation. What is the odds ratio, and its 95% confidence interval, for the effect of a motorized belt after adjusting for weight ? (2)

5.7 What is the odds ratio, and its 95% confidence interval, for a 1000 lb change in the weight of the vehicle ? (2)

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

5.9 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 77 marks on the exam: I will multiply the number of marks by 1.3 to give you a percentage.