In this practical we will be working on analysis of variance (ANOVA), both for models with continuous predictors and models with categorical predictors. The aims of the practical are to familiarise you with fitting and interpreting results from these models in Stata and also to strengthen understanding of the use of F-tests for inference.

**Data**

The data for this two session are in two files. The first includes data on children’s length, weight, age and sex from a survey carried out in rural Gambia (this dataset was used in lectures 1-3). These data are in the file "growgam1.dta".

The second dataset for the optional material is called "larvae.dta". These data refer to four experiments carried out to assess the effectiveness of carbon tetrachloride (CCl4) as a worm killer. In each experiment five rats were infested with larvae. Eight days later the rats were treated with CCl4.  After a further two days the rats were culled and the larvae counted.

The datasets can be found in folder "Linear regression datasets" on the Moodle page.

We suggest you copy them into your personal drive for use in the practical.

# Explore the data

1) Examine the dataset. How many observations are there? How many variables? Do any variables have missing values?

2) Look at a summary of values for each of the variables to get a feel for the data. Are there any implausible values?

# Compare lengths

Compare the lengths of boys and girls using:

i) a table

ii) a box plot

Which do you find more useful, the table or the box plot?

Does there appear to be a difference between girls and boys in the following statistics:

a) mean length

b) median length

c) standard deviation of lengths?

# Regression model – length by sex

We're now going to use a regression model to compare length between the sexes. As sex is a categorical variable this is done by including a dummy variable (0/1 variable) in the model.

1) Generate a new variable called "female" which takes the value 0 for boys and 1 for girls.

2) Fit a regression model relating children’s length to sex using the dummy variable you just created.

3) If we include the variable 'sex' in the model rather than the dummy variable 'female' the intercept increases from 77.8 to 79.7 but the slope stays the same. Why does this happen?

4) Complete the sentences describing the results from this model by replacing X with one of these options (0, 0, boys, girls, boys, 1): The intercept is the mean length when the dummy variable is X. The mean length of X is X cm. The slope for the dummy variable shows the change in mean length when the dummy variable is X compared to X. This means that X are 1.93cm shorter than X

5) List the assumptions that were made when fitting the model to relate sex of a child to their length. Given that the explanatory variable is categorical, is the assumption of linearity an issue? Are there any assumptions you have not checked?

# ANOVA

The comparison of length between girls and boys can also be carried out with the oneway command. Try

oneway len sex

oneway len female

1) Identify the *F*-statistic from the regress command and the oneway command. Are these the same?

2) Interpret the results of the *F*-test.

3) Using a pen, paper and a calculator, calculate the F statistic using the entries in the ANOVA table. Compare your result with the answer from Stata which you identified on the previous page.

4) Write a short paragraph to describe your findings on how weight depends on length. Discuss

# ANCOVA

The final optional part of this practical uses the larvae data.

1) Load the data into Stata and examine its contents. (Note that because there are so few observations and variables it is possible to use the list command to view the entire contents of the dataset.)

2) Produce a plot showing the distribution of the larvae counts by group.

3) Use the following regression and oneway commands to compare larvae counts between the groups. Compare the output. How do the parameter estimates relate to the group means?

regress lcount i.group

oneway lcount group

You can display the group means using the following command:

bysort group: sum lc