The data for this session are again from the study in Gambia investigating the lengths of young children. We will use a different subset of the data than the previous session. The dataset for this practical is called growgam2.dta.

The aim of the analysis here is to investigate the dependence of the child’s length on its age, sex, and mother’s height. The variables in the dataset are as follows:

sex Child’s sex (1=male; 2=female)

age Age of the child (months)

length Length of the child (cm)

mheight Height of the mother (cm)

The aim is to strengthen understanding of interpretation of results from multiple linear regression models, including why estimates change when we add additional variables.

# Explore the data

Load the data and check for any implausible values.

The main outcome variable for this analysis is the length of the child. You should therefore examine the distribution of this variable, and do an initial investigation of associations with the main exposure variables.

1) Make a plot to show the distribution of the children's lengths.

2) Plot the children’s length against

i) their age

ii) their mothers’ height

Calculate the relevant correlation coefficients in both cases.

3) Tabulate the mean and standard deviation of the children’s length by sex.

4) Make a notes of any patterns you observe.

# Univariable analysis

1) Perform univariable regression of the children’s length on:

i) the age of the child

ii) the sex of the child

iii) the height of their mother

2) Discuss the interpretation of the results. Does a child's length seem to depend on age, sex or the height of their mother?

# Multivariable analysis

1) Regress length on sex and age.

2) Discuss the intepretation of the coefficient for sex in this model. Do your results suggest there is a difference in length between boys and girls of the same age?

3) Compare the coefficient for sex from this model to the one from the univariable model including only sex as a predictor for length. Discuss how the interpretation differs between these two models.

4) Compare the mean ages of the girls and boys. Which are older?

5) Discuss why the age difference explains how the estimated mean difference in length between boys and girls changes after adjusting for age.

6) Write some notes to summarise your explantion of the differences between the findings with and without adjustment for age.

7) Compare the 95% confidence intervals for the difference in lengths between boys and girls with and without adjustment for age. Which model has narrower confidence intervals?

8) With students sat nearby, discuss possible reasons why the confidence intervals differed betwen the models. Make a note of any suggestions you have to explain this.

# Multivariable analysis 2

1) Regress length on mother’s height and age.

2) Compare your results to the univariable models including only mother's height or age as predictors. Discuss how the estimates have changed.

3) On the basis of these results what direction of association do you expect to see between mother's height and age of their child?

4) Calculate the Pearson’s correlation coefficient to see if this agrees with your prediction.

# Multivariable analysis 3

1) Regress length on gender, age and maternal height.

2) Compare each of the estimated regression coefficients with those found in earlier models.

3) In your group discuss how the results compare. What are your conclusions from this analysis?

# Assumption checks

1) For the full model containing age, sex and mother's height, use graphical methods to investigate the following assumptions:

* homoscedasticity (constant variance of the residuals);
* linear relationship between length and each predictor, conditional on the other predictors in the model;
* residuals follow a normal distribution.

You'll first need to use the predict command to get the residuals from the model, before making your plots.

2) Do your dignostic plots suggest any material violations of these assumptions?