1 Introduction to Multiple Linear Regression

- In your future studies, you will come across multiple linear regression (MLR). This is a linear model uses multiple independent variables to explain a single dependent variable.
- The implementation is very similar to simple linear regression (SLR). All that is required is to specify the additional independent variables.

```
# SLR: y explained by predictor x
Fit.slr <- lm(y~x)

# MLR: y explained by predictors x and z
Fit.mlr <- lm(y~x+z)
```

- For this case of two predictor variables, a linear relationship can be defined by the regression model

$$y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$$

.

- Again, we determine the regression coefficients, i.e. estimates for slopes and intercept. (N.B. There are variations on this notation).
  - b_0 : the intercept estimate.
  - b_1 : the slope estimate for x
  - b_2 : the slope estimate for z
- In many project datasets it is possible to implement a MLR model. For the moment, we will just look at slope and intercept estimates, their p-values and the coefficient of determination.

Simple Example

- Let try this out using the ***iris*** data set.
- We will construct a regression model using Sepal lenght as a response variable with Sepal Width and Petal Width as the predictor variables.
- *(This is not be a useful statistical analysis in practice. However we are focussing on the mechanics, so we shall proceed nonetheless).*

```
lm(Sepal.Length ~ Sepal.Width + Petal.Width,
   data=iris)
```

1.1 Model Selection

There are many important methodologies for determining which combination of predictor variables bests describes a response variable. You will meet this in future modules. We will use two simple ones for this module only.

- Adjusted Rsquared value
- The Akaike Information Criterion (AIC)

The Adjusted R-square value

The adjusted R-square value is found on the summary output for a fitted model. It is called ***adjusted*** because it takes into account the number of predictor variables being used. The law of parsimony states the simplest model that adequately explains the outcomes is the best. The candidate model with the higher adjusted R squared is considered preferable.

The Akaike Information Criterion

The AIC is a model selection metric often used in statistics. It is computed using the R command `AIC()`. The candidate model with the smallest AIC value is considered preferable.

```
fitA = lm(Sepal.Length ~ Sepal.Width + Petal.Width)
fitB = lm(Sepal.Length ~ Sepal.Width + Petal.Length)

summary(fitA)$adj.r.squared
summary(fitB)$adj.r.squared

AIC(fitA)
AIC(fitB)
```