0.1 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 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 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)
```

0.2 R square

R^2 is a measure of variation explained by regression.

The following coefficient has a natural interpretation as amount of variability in the data that is explained by the regression fit: $R^2 = SSLR/SST = 1 - SSR/SST$.

A similar interpretation is given to the adjusted coefficient R^2_{adj} which is given by $R^2_{adj} = 1 - MSR/MST$; where MSR is the mean squared error due to residuals, and MST is the total mean squared error.

The adjusted coefficient is accounting for the degrees of freedom used for each source of variation and is often a more reliable indicator of variability than R^2. R^2_{adj} is always smaller than R^2.

0.3 The Coefficient of Determination

The coefficient of determination R^2 is the proportion of variability in a data set that is accounted for by the linear model. Equivalently R^2 provides a measure of how well future outcomes are likely to be predicted by the model. (For simple linear regression, it canbe computed by squaring the correlation coefficient.)

```
summary(fit1)$r.squared
```

0.4 R square

The model with the highest R2 and adjusted R2 is the preferable of all candidate models The quadratic model is the preferable model in that case.

0.5 R square

R^2 is a measure of variation explained by regression.

The following coefficient has a natural interpretation as amount of variability in the data that is explained by the regression fit: $R^2 = SSLR/SST = 1 - SSR/SST$.

A similar interpretation is given to the adjusted coefficient R^2_{adj} which is given by $R^2_{adj} = 1 - MSR/MST$; where MSR is the mean squared error due to residuals, and MST is the total mean squared error.

The adjusted coefficient is accounting for the degrees of freedom used for each source of variation and is often a more reliable indicator of variability than R^2. R^2_{adj} is always smaller than R^2.

0.6 Adjusted R square

In a multiple linear regression model, adjusted R square measures the proportion of the variation in the dependent variable accounted for by the independent variables.