Foundations of Medical Statistics

Regression

Contents

Regression 1: Simple Linear Regression

1.1 Objectives

By the end of this session students will be able to:

- Explain, in general, the rationale behind parametric statistical models.
- Fit and interpret a simple linear regression model using a statistical package.

1.2 Recommended text books

The four recommended text books for the Foundations module as a whole (Altman; Armitage, Berry and Matthews; Van Belle, Fisher, Heagerty and Lumley; Kirkwood and Sterne) all have useful material on linear regression.

1.3 Introduction to parametric statistical models

A parametric statistical model is an algebraic description of the way in which one or more so-called **dependent** variables are influenced by **predictor** variables. Such models have wide application in medical statistics. Some examples of questions that we might wish to investigate are:

i) Does weight increase with fat intake?
ii) Can we predict the adult height of a child?
iii) Do changes in diet affect cholesterol levels when other factors are held constant?

Exercise: For each of the above, identify the dependent and predictor variables.

All non-trivial statistical models contain the following:

i) Random variables.
ii) Population parameters.
iii) Representation of uncertainty.

The dependent variables in statistical models are always random variables. Predictor variables need not be. In the relatively simple models considered in the Foundations module the predictor and dependent variables are always **observed** in a sample taken from the population, although this is not the case for all statistical models. The population parameters are unknown quantities that we wish to estimate from our sample. The uncertainty relates to variability in the dependent variable that is not explained by the predictor variables.

A statistical model makes assumptions about the form of the relationships between the dependent and the predictor variables. It should be understood at the outset that although we can examine our data to investigate the validity of a statistical model (see Regression 6 and Analytical Techniques 6), we can never be certain that it is correct.

As in many areas in statistics, terminology can cause confusion. Statisticians refer to a model with a single dependent variable as a **univariate** model and one with more than one dependent

variable as a **multivariate** model. In the Foundations module we will only be concerned with univariate models. Univariate models with a single predictor variable are termed **univariable** whilst those with more than one predictor variable are termed **multivariable**. In linear regression the univariate univariable model is often referred to as the **simple linear regression** model, whilst the univariate multivariable model is often referred to as the **multiple linear regression** model. Unsurprisingly this terminology causes confusion and it is extremely common to see univariate multivariable regression models incorrectly termed multivariate models in the medical literature.

In the statistical and medical literature predictor variables are often termed **independent** variables, to distinguish them from the dependent (or outcome) variables. This can be confusing, because this does not mean variables referred to as **independent** in this sense (predictors in a linear regression model) are independent in the usual statistical sense. It may be the case that the distribution of one predictor variable does depend upon another. To avoid confusion we will not use the term 'independent' in this way in this module. Other terms that are used in the statistical and epidemiological literature and which will be used in this module are **explanatory** (synonymous with **predictor**) and **response** and **outcome** (both synonymous with dependent).

A further question of terminology concerns the use of the term 'model'. Our use here designates what in other contexts is called the **linear predictor**: this describes the algebraic relationship between the mean of the dependent variable and the predictor variables. However, there is another use of 'model' that we have also met in the Foundations module: **probability model**. This describes how the random element of the dependent variable is generated. For linear regression models the assumption is of a normal probability model. Next term, in Generalised Linear Models, other probability models are considered, and there the term linear predictor will be used to refer specifically to the algebraic relationship.

1.4 Simple linear regression model

The simple linear regression model relates a single predictor variable to a single dependent variable. We will use the following two examples to illustrate this model.

1.4.1 Examples

Example 1: The scatter plot in Figure 1 shows data on weight and age from a cross-sectional study of children living in a rural area in The Gambia.

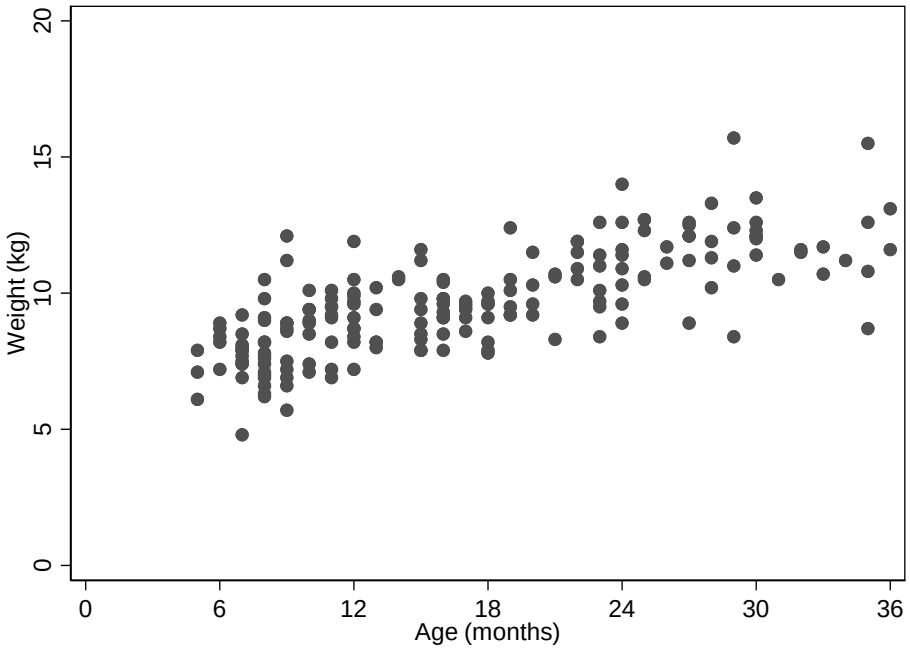


Figure 1: Age and weight of children in the Gambian cross-sectional survey.

Example 2: Table 1 shows the age (in months) of 11 infants at the time when they started to walk alone. These infants, as newborns, were randomised to two groups, an eight week active exercise group and an eight week control group. This would usually be analysed with a two sample t-test (see Analytical Techniques 5), but later on in this session we will show that it can also be analysed with a simple linear regression model.

Age in months for walking alone	
Active exercise group (n=6)	Eight week control group (n=5)
9.00, 9.50, 9.75, 10.00, 13.00, 9.50	13.25, 11.50, 12.00, 13.50, 11.50

Table 1: Children's ages at time of first walking alone by randomisation group

1.4.2 The distinction between dependent and predictor variables

In investigating the association (Analytical Techniques 4) between two variables (*X* and *Y*) no distinction is made between the roles of *X* and *Y*. The extent of the association between *X* and *Y* is the same as that between *Y* and *X*. Before defining a regression model we have to decide which of *Y* and *X* is the dependent variable. Sometimes this is obvious from the context. In example 2 it is natural to consider the walking age as the dependent variable: conceptually it would be misguided to investigate the extent to which randomisation group is predicted by age at walking. However in example 1 it is possible that we might be interested in age as a predictor of weight or weight as a predictor of age. Let us assume that the primary focus is on age as a predictor of weight.

1.4.3 The mean function

Having determined that the main focus in Example 1 is on the dependency of weight (Y) on age (X) we have to propose a form for the relationship between the distribution of Y and X. The first key component of this is the **mean** function. This is defined to be:

$\mathrm{E}(Y|X = x)$, the "expected value of Y when X takes the value x"

In general the mean function can be any function of X. In the simple linear regression model the mean function is assumed to be linear.

$$\mathrm{E}(Y \mid X = x) = \alpha + \beta x$$

This specification of the mean function has two parameters, α and β.

$\boldsymbol{\alpha}$ is the **intercept**. It is the expectation of Y when X takes the value 0.

$\boldsymbol{\beta}$ is the **slope**. It is the increase in the expectation of Y per one unit increase in X.

It is worth emphasising that this is an assumed model for the relationship between the two variables. It is possible that the relationship takes a more complex form. Figure 2 illustrates a linear mean function and a more complex one (obtained using a cubic spline function). By eye, the fit of the linear mean function seems reasonable (although perhaps there is a suggestion of curvature, with a steeper slope at younger ages). The spline function may 'follow the data' better, but there is a suspicion that the continuously changing gradient may be an artefact of the sample, rather than representing the relationship in the population.

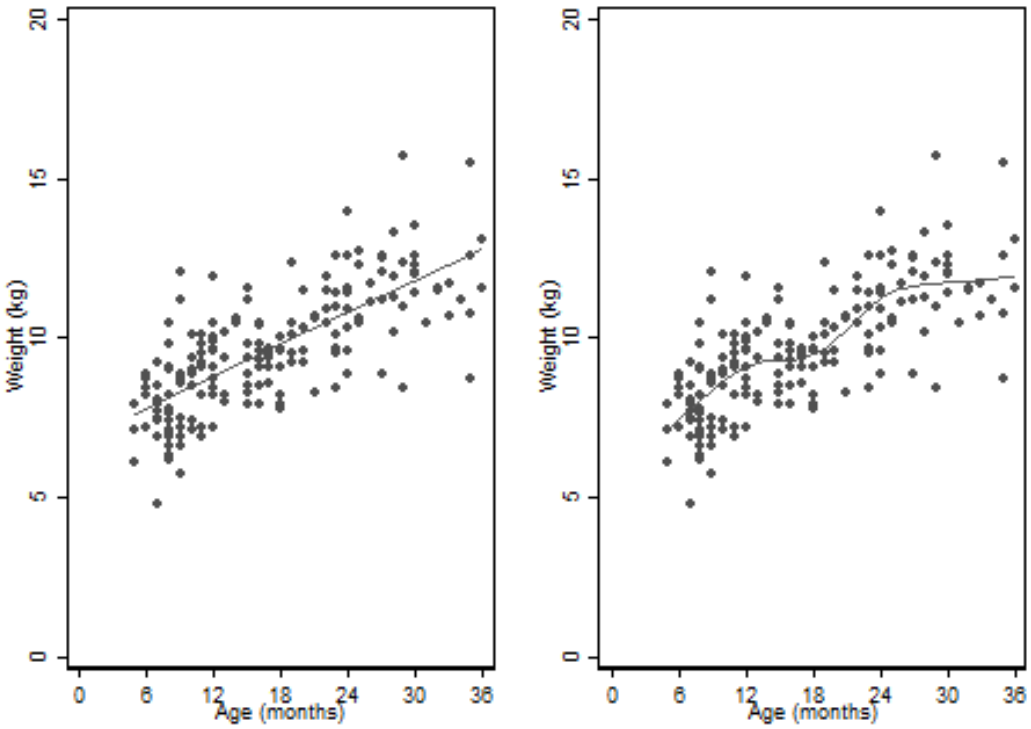


Figure 2: Possible forms for the mean function for the data from the Gambian cross-sectional survey.

1.4.4 The conditional distribution and the variance function

In order to fully specify a statistical model it is necessary to describe the distribution of the dependent variable conditional on the predictor variable. The simple linear regression model is sometimes considered to be defined purely by the assumption that the **variance** of the dependent variable (conditional on the predictor variable) is constant. However, in order to allow statistical inference we further assume that the conditional distribution is a **normal distribution** (with constant standard deviation σ). This is a simple assumption to make, but as with assumptions made about the mean function, it may or may not be reasonable. For the data from The Gambia the assumption seems a reasonable one: at all ages the spread of points is not obviously skewed, nor does the spread of points appear to change markedly with age. Figure 3 illustrates some hypothetical data for which this assumption would not be reasonable.

Exercise: What aspect of the data in Figure 3 suggests that a simple linear regression model is not appropriate?

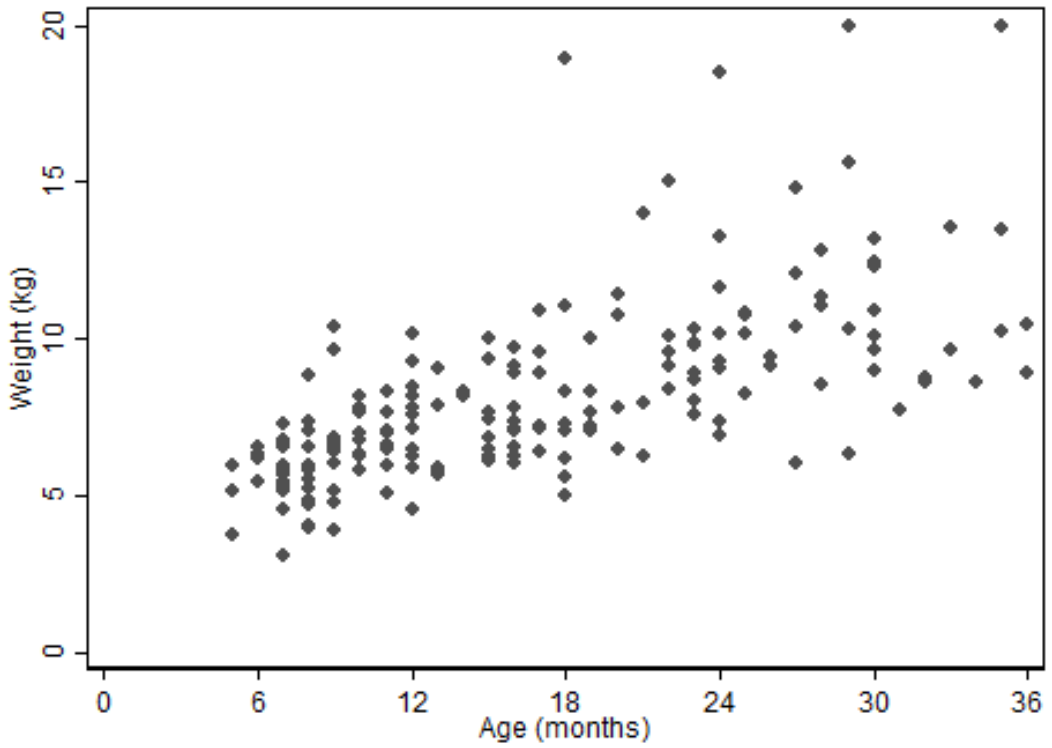


Figure 3: Hypothetical data for which a simple linear regression model would strictly be inappropriate.

The variance σ^2 is termed the **residual variance** in linear regression models.

1.4.5 The simple linear regression model: definition

The simple linear regression model relating one random variable (Y) to another (X) is:

$$(Y \mid X = x) \sim N(\alpha + \beta x, \sigma^2)$$

This definition makes explicit the underlying assumed probability model for *Y* conditional on *X*. The same model can also be given as:

$$y = \alpha + \beta x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

If we have a sample of size *n* and denote the realisations of *Y* and *X* by Y_i and X_i (i = 1,..,n) respectively, we can write this model as:

$$(Y_i | X_i = x_i) \sim NID(\alpha + \beta x_i, \sigma^2) \quad i = 1, \ldots, n \tag{1}$$

Here *NID* stands for '**Normally and Independently Distributed**'. A key assumption of the simple linear regression model is that all of the observations are independent. For the data from the Gambia this assumption is likely to be reasonable if each of the children comes from a different family. If the data included siblings then the assumption of independence would not be a reasonable one. Methods of analysing data that are not independent are introduced in the module 'Analysis of hierarchical and other dependent data' in Term 2.

Equation (1) is often re-written as

$$(Y_i | X_i = x_i) = \alpha + \beta x_i + \varepsilon_i, where\ \varepsilon_i \sim NID(0, \sigma^2) \tag{2}$$

or more succinctly (but less explicitly) as

$$y_i = \alpha + \beta x_i + \varepsilon_i, where\ \varepsilon_i \sim NID(0, \sigma^2) \tag{3}$$

This relationship can also be expressed using matrix algebra.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad where\ \varepsilon \sim N(0, \boldsymbol{I}\sigma^2) \tag{4}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & . \\ 1 & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

In this formulation $\boldsymbol{X}$ is an $n \times 2$ matrix, $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are vectors of length *n* whilst $\boldsymbol{\beta}$ is a vector of length 2 whose first element is α (sometimes written as β_0). These are the most commonly adopted ways of writing these equations.

The residuals are assumed to follow a **multivariate normal distribution** with variance-covariance matrix equal to σ^2 multiplied by the identity matrix. This is equivalent to assuming that residuals are normally and independently (implying that the covariances in the multivariate normal formulation are all zero) distributed with constant variance σ^2.

Exercise: Write down the linear regression model relating weight to age for our Gambian survey example. What is the interpretation of α, β and σ?

1.4.6 Residuals

Equations (2) and (3) include a representation of the error term (ε_i) in the simple linear regression model:

$$\varepsilon_i = y_i - (\alpha + \beta x_i)$$

ε_i is the **random error** or (**true**) **residual** for the *i*th observation. It is the difference between the observed value (y_i) and its value as predicted from the model ($\alpha + \beta x_i$). Note that these true residuals are defined in terms of deviations from the model defined by population parameters (α and β). The term residual is also used to define deviations from a fitted regression model (*i.e.* a model in which α and β are replaced by estimates from a sample, $\hat{\alpha}$ and $\hat{\beta}$). The terms **true residual** and **observed (or fitted) residual** can be used to make this distinction clear.

1.5 Estimation of parameters

In the specification of the simple linear regression model there are three population parameters (α, β and σ). The object of a statistical analysis is to make inferences about these population parameters from an observed sample of Y_i and X_i (i = 1,..,n). There are many different methods available. In this section we focus on one approach termed **ordinary least squares**. We also comment on the relationship between estimates obtained using this method and those obtained by maximum likelihood.

As throughout the Foundations module we use "hat" to denote estimates, *e.g.* $\hat{\beta}$ is the estimated value of β. The **fitted** value for the *i*th observation is the **estimated expectation** of Y conditional on $X = x_i$. It is denoted by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

The **observed (or fitted or estimated) residual** for the *i*th observation (usually referred to simply as 'the residual') is denoted by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

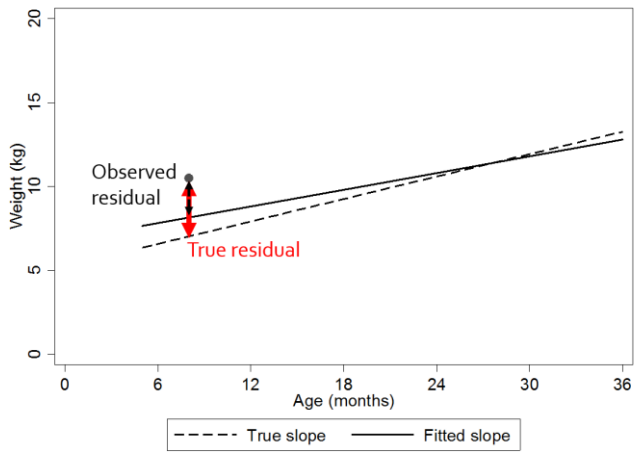


Figure 4: Illustration of a residual for a single realisation of Y at a particular value of X.

1.5.1 Ordinary Least Squares (OLS) estimates of α and β

The OLS estimators of α and β minimise the sum of squared deviations from the fitted regression line. Formally the OLS estimators are the values of $\hat{\alpha}$ and $\hat{\beta}$ that minimise the **residual sum of squares**, given by:

$$SS_{RES} = \sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 \tag{5}$$

The ordinary least squares estimates of α and β are given by the following.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{6}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

$$where\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}\ \ and\ \ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Proof:

To solve for the value of α that minimises the SS, we differentiate (5) with respect to $\hat{\alpha}$ and set the differential to zero:

$$\frac{d(SS_{RES})}{d(\hat{\alpha})} = \sum_{i=1}^{n} -2\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0$$

Since $\sum_{i=1}^{n}(y_i) = n\bar{y}$ and $\sum_{i=1}^{n}(x_i) = n\bar{x}$ (see above) we can simplify to:

$$-n\bar{y} + n\hat{\alpha} + n\hat{\beta}\bar{x} = 0$$

Rearranging the above and divide by *n* to give:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

To solve for the value of β that minimises the SS we have to differentiate (5) with respect to $\hat{\beta}$. First we substitute in our solution for $\hat{\alpha}$ into (5) as follows:

$$SS_{RES} = \sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \sum_{i=1}^{n} \left(y_i - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta} x_i\right)^2 = \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right)^2 \tag{8}$$

Now differentiating (8) with respect to $\hat{\beta}$ and setting the differential to zero gives

$$\frac{d(SS_{RES})}{d(\hat{\beta})} = \sum_{i=1}^{n} -2(x_i - \bar{x})(y_i - \bar{y}) + 2\hat{\beta}(x_i - \bar{x})^2 = 0$$

Rearranging gives:

$$\hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The estimates in equations (6) and (7) are also the **maximum likelihood estimates** of α and β.

Proof:

If $Y_i \sim NID(\mu,\sigma^2)$, the log likelihood function is:

$$l(\mu \mid y_1 \ldots y_n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

So, for the simple linear regression model:

$$l(\alpha, \beta \mid y_1 \ldots y_n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2$$

Therefore, for any fixed positive value for σ^2, maximising the log likelihood function is equivalent to minimising (5) and so the OLS estimates are also maximum likelihood estimates of α and β.

1.5.2 Estimation of the residual variance (σ^2)

Since the residual variance is the mean squared size of the residuals, an intuitively appealing estimator of σ^2 is given by dividing the residual sum of squares by the number of observations, *i.e.*

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^{\,2}}{n} = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 / n \qquad (9)$$

It can be shown that this is the maximum likelihood estimator of σ^2. However it can also be shown that the estimator in (9) is not an unbiased estimator of σ^2. The bias arises because the

observed values tend, on average, to lie closer to the fitted line (defined by $\hat{\alpha}$ and $\hat{\beta}$) than they do to the true regression line (defined by α and β). This is an exact parallel to the way the variability of a sample around its mean underestimates the variability around the population mean (seen in the Inference course).

This point is seen most clearly by considering fitting a simple regression line to a sample of size two. Provided that σ^2 is not equal to zero then neither point will lie exactly on the true regression line, but both of them will lie on the fitted line and the estimate of σ^2 from equation (9) will be zero.

It can be shown that an unbiased estimator of the residual variance in the simple linear regression model is given by:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^{\,2}}{(n-2)} = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 / (n-2) \qquad (10)$$

This quantity is referred to as the **residual mean square**. It is often denoted by $MS_{RES.}$ The denominator is $(n-2)$ because fitting the model first requires the estimation of two parameters (α and β). The estimation of these parameters is said to reduce the information about the variance by **two degrees of freedom**. Recall from Inference that this conforms with the general rule that the degrees of freedom are reduced by the number of estimated parameters.

Exercise: By hand, calculate estimates of α, β and σ^2 for the following sample (X, Y) of size 4: (1, 6), (2, 8), (4, 10), (5, 12).

1.6 Simple linear regression in Stata

Linear regression models can be fitted in many different ways in Stata. The simplest command is **`regress`**. The basic syntax is:

```
. regress <depvar> <predvars>
```

where **`depvar`** is the dependent variable and **`<predvars>`** is a list of predictor variables.

1.7 Examples

1.7.1 Example 1: Cross-sectional study of Gambian children using Stata

```
. regress wt age

      Source |       SS       df       MS       Number of obs =     190
-------------+------------------------------    F(  1,   188) =  221.39
       Model |  359.063204     1  359.063204    Prob > F      =  0.0000
    Residual |  304.906554   188  1.62184337    R-squared     =  0.5408
-------------+------------------------------    Adj R-squared =  0.5383
       Total |  663.969759   189  3.51306751    Root MSE      =  1.2735

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1653314   .0111115    14.88   0.000     .1434121    .1872507
       _cons |   6.837584   .2100701    32.55   0.000     6.423187    7.251982
------------------------------------------------------------------------------
```

In Regression 2 and Regression 3 all aspects of this output will be explained. Here we focus on the estimates of α, β and σ^2, each of which is highlighted in **bold**.

$\hat{\alpha} = 6.84$. The estimated mean weight of a child aged 0 months is 6.84 kg. Since there are no children aged 0 months in the study this is an **extrapolation** based on weights of children aged between 6 and 36 months and an assumption of linearity. It should therefore be interpreted cautiously.

$\hat{\beta} = 0.165$. The mean weight of children is estimated to increase by 0.165 kg per month. From $\hat{\alpha}$ and $\hat{\beta}$ we can estimate the mean weight at any age. Figure 5 shows these fitted values.

$\hat{\sigma}^2 = 1.62$ (and $\hat{\sigma} = 1.27$). Points are scattered around the line with a standard deviation of 1.27 kg.

Exercise: Using the estimates of α, and β, calculate the estimated weight for an infant aged 12 months.

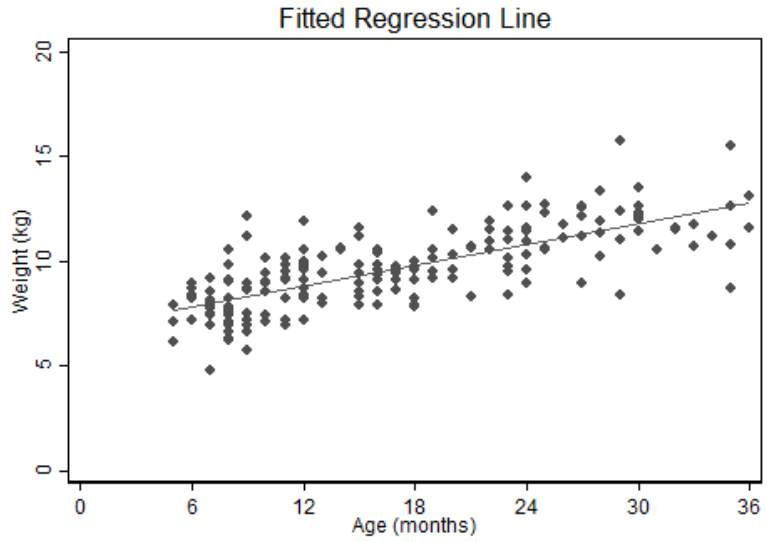


Figure 5: Data and fitted values for the Gambian cross-sectional survey.

1.7.2 Example 2: Randomised trial comparing exercise regimes in babies

To use the linear regression model to analyse this data it is necessary to first create a 'dummy' variable (**rand** in the Stata output below) that takes the value 0 in the active exercise group and 1 in the control group. Figure 6 shows the walking ages plotted against this dummy variable (the variable is termed **rand** in the Stata output).

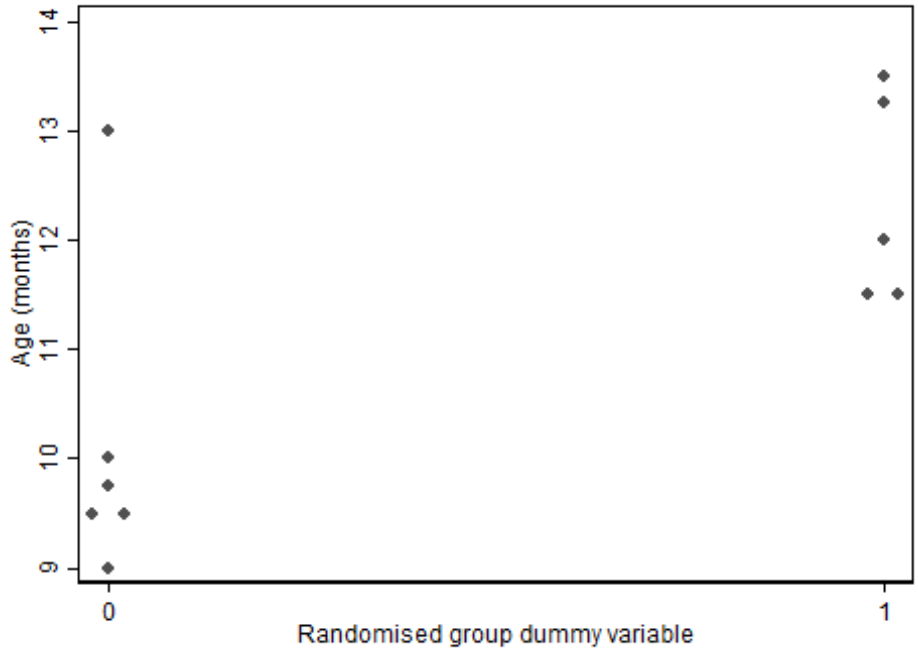


Figure 6: Data for the exercise regime trial: age at walking by randomisation group. Points representing equal walking ages have been 'jittered' (moved a small distance to the left or right) to aid clarity. The 'dummy' variable takes the value 0 in the active exercise group and 1 in the control group.

If we now fit a linear regression model to these data treating the 'dummy' variable as the predictor variable and age at walking as the outcome variable we obtain the following results

```
. regress age rand

      Source |       SS       df       MS        Number of obs =      11
-------------+------------------------------     F(  1,     9) =    8.58
       Model |  13.5017045     1  13.5017045     Prob > F      =  0.0168
    Residual |    14.16875     9  1.57430556     R-squared     =  0.4879
-------------+------------------------------     Adj R-squared =  0.4311
       Total |  27.6704545    10  2.76704545     Root MSE      =  1.2547

------------------------------------------------------------------------
      age |   Coef.  Std. Err.    t    P>|t|  [95% Conf. Interval]
----------+-------------------------------------------------------------
     rand |    2.225  .7597667   2.93  0.017  .5062884    3.943712
    _cons |  10.125  .5122346  19.77  0.000  8.966245    11.28376
------------------------------------------------------------------------
```

$\hat{\alpha} = 10.125$. This is interpreted as the estimated mean walking age in months of a child with 'dummy' variable equal to 0 *i.e.* it is the estimated mean age at walking in the active exercise group.

$\hat{\beta} = 2.225$. The mean age at walking is estimated to increase by 2.225 months per unit increase in the 'dummy' variable. **However since a unit increase in the dummy variable equates to moving from the active exercise group to the control group we can interpret this as the difference in mean walking ages between the groups**. This is illustrated in Figure 7 which shows the line of fitted values from this model. The difference in the means is 2.225 months (and hence the mean in the control group is 10.125 + 2.225 = 12.35 months).

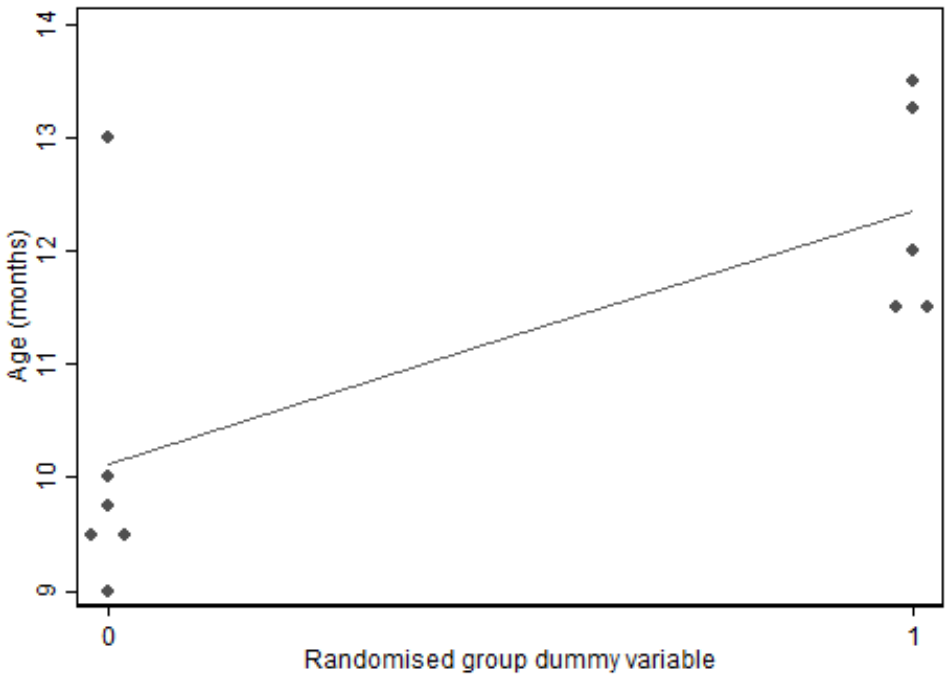


Figure 7: Exercise regime trial: age at walking by randomisation group. Data points and fitted values.

What has effectively been done here is to make a judicious choice when coding the predictor variable, so that a 1 unit increase in the predictor variable equates to the difference between groups. This means that the slope of the line can be interpreted as the difference in the mean outcome variable between the two groups. We return to this important topic in Regression 3 and later sessions.

In the Stata analysis above the **`rand`** variable was defined to take the values 0 and 1 in the two groups. However, the **`i.`** prefix in the **`regress`** command in Stata can be used to create the necessary dummy variables for analysis without us needing to do the recoding. This tells Stata that any predictor variables written as **`i.<var>`** are categorical predictor variables that should be converted to dummy variables before they are included in the model.

The result of using the regress command with **i** prefix is illustrated below using the variable **group** which was coded as 1 in the active exercise group and 2 in the control group.

```
. regress age i.group

      Source |       SS       df       MS          Number of obs =      11
-------------+------------------------------       F(  1,     9) =    8.58
       Model |  13.5017045     1  13.5017045        Prob > F      =  0.0168
    Residual |    14.16875     9  1.57430556        R-squared     =  0.4879
-------------+------------------------------       Adj R-squared =  0.4311
       Total |  27.6704545    10  2.76704545        Root MSE      =  1.2547

------------------------------------------------------------------------------
         age |   Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |    2.225   .7597667     2.93   0.017     .5062884    3.943712
       _cons |   10.125   .5122346    19.77   0.000     8.966245    11.28376
------------------------------------------------------------------------------
```

Regression 2: Properties of Ordinary Least Squares Estimators and Inference

2.1 Objectives

By the end of this session students will be able to:

- Describe the main properties of ordinary least squares estimators in the simple linear regression model.
- Explain the relationships between parameters and test statistics in the simple linear regression model and the Pearson correlation coefficient.

2.2 Introduction

In Regression 1 the simple linear regression model was introduced and the following estimators given for the parameters.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

(1) [eq (6) in regression 1]

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(2) [eq (7) in regression 1]

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^{\,2}}{(n-2)} = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 / (n-2)$$

(3) [eq (10) in regression 1]

In this session the properties of these estimators will be explored and links between simple linear regression and association as measured by the Pearson correlation coefficient (Analytical Techniques 4) explained. The following notation and terminology will be used.

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ "Sample mean of *Y*"

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ "Sample mean of *X*"

$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ "(Corrected) sum of squares of *Y*"

$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ "(Corrected) sum of squares of *X*"

$SD_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{SS_{yy}}{n-1}$ "Sample variance of *Y*"

$SD_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{SS_{xx}}{n-1}$ "Sample variance of *X*"

$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ "Sum of cross-products"

$CV_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{S_{xy}}{n-1}$ "Sample covariance"

$r_{xy} = \frac{CV_{xy}}{SD_x SD_y}$ "Sample (Pearson) correlation"

$SS_{RES} = \sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$ "Residual sum of squares"

2.3 Some properties of the OLS estimators

i) The estimated regression line passes through the centre of the data.

Proof:

From (1) $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
and fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$

It follows that $\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ (4)
So, $\hat{y}_i = \bar{y}$ when $x_i = \bar{x}$

ii) Assuming that the model is correct, $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}^2$ are each unbiased estimators (*i.e.* $E(\hat{\alpha}) = \alpha$ *etc.*).
iii) $\hat{\alpha}$ and $\hat{\beta}$ (but not $\hat{\sigma}^2$) are maximum likelihood estimators (see Regression 1).
iv) $\hat{\alpha}$ and $\hat{\beta}$ are **efficient** estimators of α and β (*i.e.* their variances are the smallest among all linear estimators). This result is known as the Gauss-Markov theorem.

Results ii) & iv) are not proved here, but can be found in standard text books.

2.4 Specific properties of $\widehat{\boldsymbol{\beta}}$

In many applications where simple linear regression is used the primary focus of the analysis is on the regression slope (β), since this is the parameter that reflects the way in which differences in the predictor variable (X) are related to differences in the dependent variable (Y).

2.4.1 Alternative methods of calculating and expressing $\widehat{\boldsymbol{\beta}}$

The parameter estimator can also be written as the ratio of the covariance between the predictor and dependent variable to the variance of the predictor variable *i.e.*

$$\hat{\beta} = \frac{S_{xy}}{SS_{xx}} = \frac{CV_{xy}}{SD_x^2} \qquad (5)$$

2.4.2 Regression of *Y* on *X* and *X* on *Y*

If we use the notation $\hat{\beta}_{y|x}$ to denote the estimate of the slope from a simple linear regression model with predictor variable X and dependent variable Y it follows from (5) that

$$\hat{\beta}_{y|x} = \frac{CV_{xy}}{SD_x^2} \qquad \text{and} \qquad \hat{\beta}_{x|y} = \frac{CV_{xy}}{SD_y^2}$$

Hence,

$$\hat{\beta}_{y|x}\hat{\beta}_{x|y} = r_{xy}^2 \qquad (6)$$

Exercise: Using the correlation coefficient and parameter estimates below verify equation (6) for the data from the Regression 1 practical.

```
. corr chol1 chol2

         ... |    chol1    chol2
       chol2 |   0.6179   1.0000

.  regress chol2 chol1
------------------------------------------------------------------------------
       chol2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       chol1 |   .5786806   .0747598     7.74   0.000     .4303031     .727058
------------------------------------------------------------------------------

. regress chol1 chol2
------------------------------------------------------------------------------
       chol1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       chol2 |   .6598362   .0852443     7.74   0.000     .4906499    .8290225
 ------------------------------------------------------------------------------
```

Equation (6) also demonstrates that the slope of the regression line relating *Y* to *X* is only equal to the reciprocal of the slope of the regression line relating *X* to *Y* if the magnitude of the correlation between the two variables is 1. This illustrates the fact that (unless the two variables are perfectly correlated) regression of *Y* on *X* yields a different regression line from regression of *X* on *Y* (figure 1).

This is an important result, often not appreciated by researchers. Its implication is that if we wish to fit a regression model in order to use one random variable to predict another then it is vital that that we include the variable that we wish to predict as the dependent variable in the model.

Example 1: Cross-sectional study of Gambian children

Using Stata to analyse this data gives the following results.

```
. regress wt age

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1653314   .0111115    14.88   0.000     .1434121    .1872507
       _cons |   6.837584   .2100701    32.55   0.000     6.423187    7.251982
------------------------------------------------------------------------------

. regress age wt

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wt |      3.2709   .2198297    14.88   0.000     2.837251     3.70455
       _cons |  -14.56803   2.159658    -6.75   0.000     -18.8283   -10.30775
------------------------------------------------------------------------------
```

The first model shows that the mean weight of children is estimated to increase by 0.165 kg per month. It might be tempting to conclude, on the basis of this, that the mean age of children is estimated to increase by 1/0.165 = 6.1 months per kg increase in weight. The results from the second model illustrate that this is incorrect. In fact the mean age of children is estimated to increase by 3.3 months per kg increase in weight. Figure 1 shows the two fitted regression lines for this data.

Also note, in passing, that the statistical significance of the slope in the two regression models is the same. This important result will be returned to in section 2.7.2.

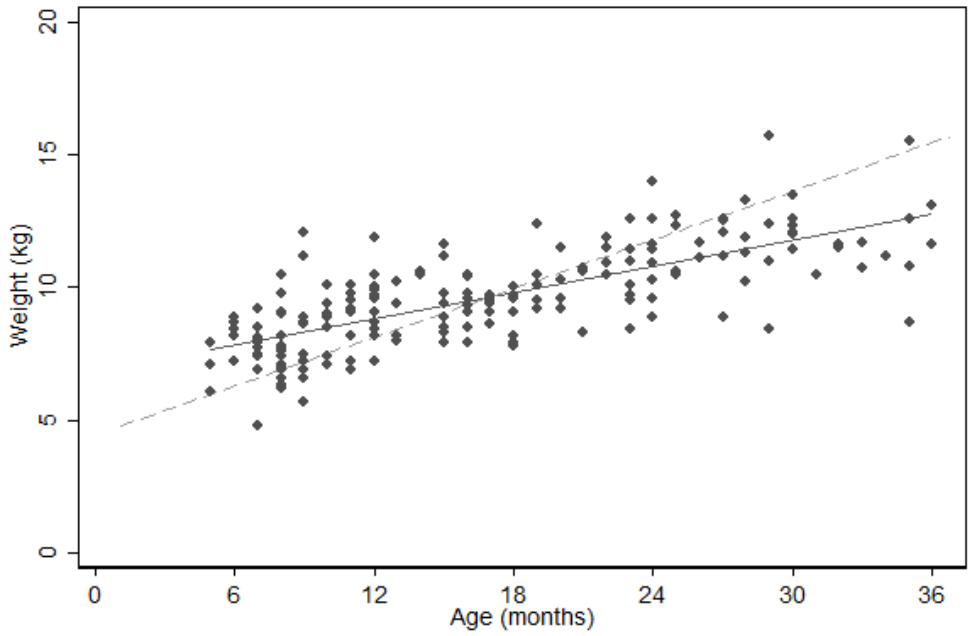


Figure 1: Data and fitted values from simple linear regression models relating weight to age (solid line) and age to weight (dashed). Data from the Gambian cross-sectional survey.

2.5 Variance and covariance of the estimators of slope and intercept

Assuming that the model is correct the variances of $\hat{\alpha}$ and $\hat{\beta}$ are as follows.

$$V(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right) = \frac{\sigma^2}{(n-1)}\left(1 - \frac{1}{n} + \frac{\bar{x}^2}{SD_x^2}\right) \qquad (7)$$

$$V(\hat{\beta}) = \frac{\sigma^2}{SS_{xx}} = \frac{\sigma^2}{(n-1)SD_x^2} \qquad (8)$$

From (7) and (8) it can be seen (as might be expected intuitively) that the variance of both estimators increases with the size of the residual variance and decreases with increasing sample size. In addition the variance of the slope decreases as the spread of *X* values (as measured by the standard deviation) increases.

The variance of the intercept increases with increasing mean. This too is an intuitive result since the line of best fit might be anticipated to be 'anchored most firmly' where there is 'most' data (*i.e.* at the mean value). The further that $X = 0$ is from the mean, the less precise is the estimate of the fitted value.

The estimators of the slope and the intercept are not, in general, independent. Their covariance is given by:

$$Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2\bar{x}}{SS_{xx}} \qquad (9)$$

Equations (7) – (9) all include the residual variance, which is unknown. If we wish to estimate the variance of the parameters and their covariance we must replace σ^2 with $\hat{\sigma}^2$.

2.5.1 Centring

A common technique when fitting a simple linear regression model is to centre the predictor variable by subtracting the mean value from all measurements before fitting the linear regression model. This doesn't alter the fitted regression line (in the sense that the predicted value for each value of the original variable is unchanged), but the change of scale changes the intercept. The intercept now has the interpretation of being the mean value of the dependent variable when the centred variable is zero *i.e.* when the original variable is at its mean level. This is illustrated in the following Stata output.

```
. summ age wt

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       190    16.97895    8.336798          5         36
          wt |       190    9.644737    1.874318        4.8       15.7

. regress wt age

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .1653314   .0111115    14.88   0.000     .1434121    .1872507
   _cons |   6.837584   .2100701    32.55   0.000     6.423187    7.251982
------------------------------------------------------------------------------

. gen age_cen=age-16.97895

. regress wt age_centred

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 age_cen |   .1653314   .0111115    14.88   0.000     .1434121    .1872507
   _cons |   9.644737   .0923906   104.39   0.000     9.462482    9.826993
------------------------------------------------------------------------------
```

Note that centring does not alter the estimated slope or its variance. Further the 'new' intercept is simply the mean weight (see 2.3 property i)). As predicted by equation (7) the variance of the new intercept is less than that from the un-centred model. Additionally for the centred analysis the estimators of slope and intercept are uncorrelated (equation (9)).

2.6 Inference

Both $\hat{\alpha}$ and $\hat{\beta}$ can be written as linear combinations of the observed values of Y and are therefore functions of the random errors (ε_i). For example

$$\hat{\beta} = \sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})}{SS_{xx}}(y_i-\bar{y})\right] = \beta + \sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})}{SS_{xx}}(\varepsilon_i-\bar{\varepsilon})\right]$$

substituting in $(y_i-\bar{y}) = \beta(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon})$

So, since the $\varepsilon_i \sim NID(0,\sigma^2)$, the estimated parameters are normally distributed because they are linear combinations of the ε_i.

2.6.1 Hypothesis testing

We can calculate a Wald statistic to test the null hypothesis H_0:β=0 against the alternative H_1:$\beta \neq 0$. The test statistic is

$$t = (\hat{\beta} - 0)/SE(\hat{\beta}) \qquad (10)$$

Where, from (8) and replacing σ^2 with $\hat{\sigma}^2$:

$$SE(\hat{\beta}) = \sqrt{V(\hat{\beta})} = \hat{\sigma}/\sqrt{SS_{xx}}$$

The replacement of σ^2 with $\hat{\sigma}^2$ means that *t* follows a *t*-distribution with (*n*-2) degrees of freedom (rather than a *z*-distribution) if H_0 is true. This permits calculation of a *p*-value to test the null hypothesis that the slope is zero.

Because the assumed distribution of the random errors is normal, the log-likelihood is quadratic and the Wald test is asymptotically equivalent to a likelihood ratio test.

Exercise: For the exercise regime randomised trial data (Stata output below) perform a test of the null hypothesis that mean age at walking is the same in both groups. Refer to Regression 1 notes to confirm that your answer is correct.

```
Number of obs =      11

------------------------------
      age |   Coef.  Std. Err.
----------+-------------------
    group |   2.225  .7597667
    _cons |  10.125  .5122346
------------------------------
```

2.6.2 Confidence intervals for α and β

The 95% confidence interval for the parameter β is in a simple linear regression model is given by:

$$\hat{\beta} \pm t_{n-2,\,0.975} SE(\hat{\beta}) \qquad (11)$$

where $t_{(n\text{-}2),\,0.975}$ represents the 97.5th centile of a *t* distribution with (*n*-2) degrees of freedom.

For the Gambian cross-sectional survey data $\hat{\beta} = 0.165,\ SE(\hat{\beta}) = 0.0111$. *n* here is 190 and $t_{(n\text{-}2),\,0.975}$ =1.973. So the 95% confidence interval for β is 0.165±1.973×0.0111=(0.143,0.187) as shown in the output in 2.5.1.

An analogous approach can be used to calculate a confidence interval for the intercept in a simple linear regression model. i.e. $\hat{\alpha} \pm t_{n-2,\,0.975} SE(\hat{\alpha})$

Exercise: For the exercise regime randomised trial data construct a 95% confidence interval for the difference in mean walking age between the two groups.

2.6.3 Confidence intervals for a predicted value

Confidence intervals can also be constructed for the predicted value at a particular value of the predictor variable ($X=x$). The use of the term "predicted" value can cause confusion here as it might be taken to imply we want a confidence interval for prediction of where an individual observation might lie. In fact the predicted value is $y_x = \alpha + \beta x$ and its interpretation is an **expectation** *i.e.* E($Y|X=x$).

Its confidence interval is found by first specifying that the estimator for y_x is $\hat{y}_x = \hat{\alpha} + \hat{\beta}x$ and computing the variance of this. Equations (7) – (9) can be used to show that its variance is given by:

$$V(\hat{y}_x) = \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}\right] \qquad (12)$$

The 95% confidence interval is therefore given by:

$$\hat{y}_x \pm t_{n-2,\,0.975}\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}\right]} \qquad (13)$$

Figure 2 shows this 95% confidence interval for the cross-sectional survey data from The Gambia. The grey area around the fitted slope identifies the 95% confidence interval around each predicted value.

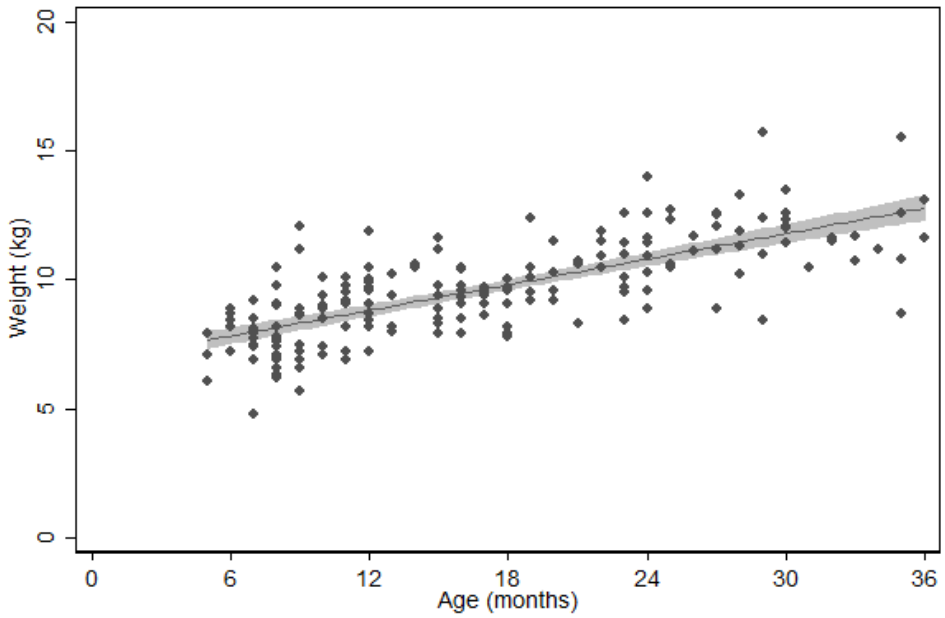


Figure 2: 95% confidence intervals for fitted values from a simple linear regression model relating weight to age. Data from the Gambian cross-sectional survey.

Predictions and 95% confidence intervals for these can be made for values of the predictor variable that occur both in the data being analysed and values not in the data. The formula and figure demonstrate that the width of the confidence interval increases with the distance from

the mean. Care must be taken when extrapolating outside the range of the observed data as an un-testable assumption of continuing linearity is being made by doing this.

2.6.4 Reference range for individual values

In some situations we may wish to estimate a reference range for individual observations having a particular value of X. A 95% reference range is the interval in which 95% of observations are expected to lie. To estimate this for $X=x$ we have to take into account both the imprecision in $E(Y|X=x)$ and the additional random error in an observation (which has variance σ^2). Hence the variance of such an individual prediction is given by:

$$V(\hat{y}_x) + \sigma^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \qquad (14)$$

Therefore the 95% reference range is the interval:

$$\hat{y}_x \pm t_{n-2,\,0.975}\hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$$

Figure 3 shows this 95% reference for the cross-sectional survey data from The Gambia. It is important to realise that this is not a confidence interval since a prediction for an individual is a random variable not a parameter. For this reason the term 'reference range', rather than 'confidence interval', is used.

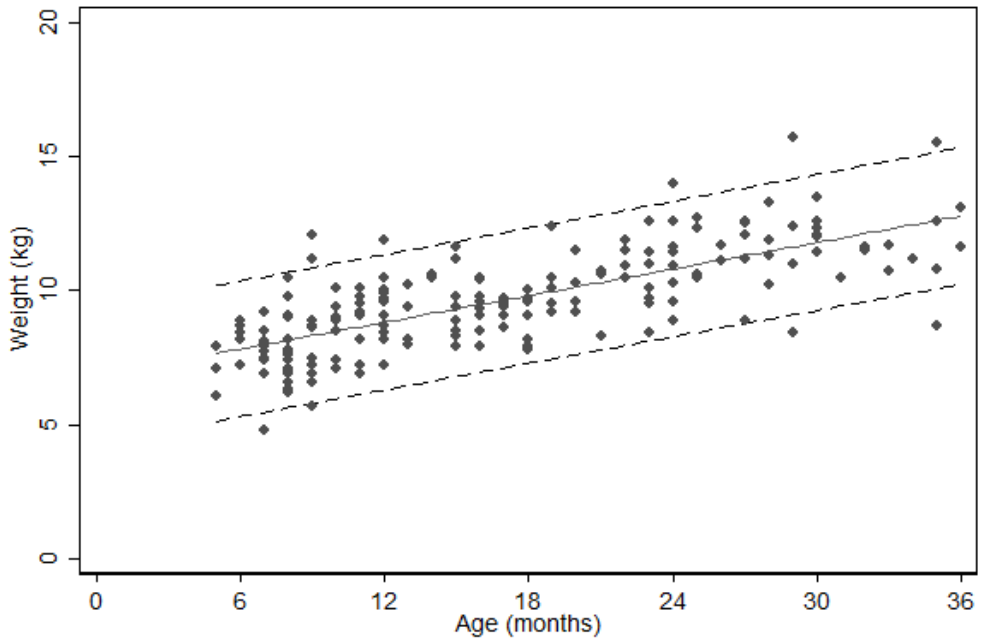


Figure 3: 95% reference range calculated from a simple linear regression model relating weight to age. Data from The Gambian cross-sectional survey.

2.7 Linear regression models and the Pearson correlation coefficient

The distinction between simple linear regression and an investigation of association between two variables has been emphasised both in Analytical Techniques 4 and Regression 1. Nonetheless there are some important relationships between the parameter estimates and test statistics introduced here and the Pearson correlation coefficient used to quantify associations. The first of these has been described in 2.4.2. Here we consider two other important results.

2.7.1 Interpretation of r^2 as the proportion of variance explained

It is shown below that there is a relationship between the Pearson correlation coefficient, the sum of squares of the dependent variable and the residual sum of squares. The relationship is:

$$r^2 = \frac{SS_{yy} - SS_{RES}}{SS_{yy}} = 1 - \frac{SS_{RES}}{SS_{yy}} \qquad (15)$$

Proof:

$$\frac{SS_{RES}}{SS_{yy}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

From (2):

$$\frac{SS_{RES}}{SS_{yy}} = \frac{\sum_{i=1}^{n}\left((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} - \frac{2\hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{2\hat{\beta}S_{xy}}{SS_{yy}} + \frac{\hat{\beta}^2 SS_{xx}}{SS_{yy}}$$

Substituting in for:

$$\hat{\beta} = \frac{S_{xy}}{SS_{xx}} \quad \text{and} \quad r^2 = \frac{S_{xy}^2}{SS_{xx}SS_{yy}}$$

Gives,

$$\frac{SS_{RES}}{SS_{yy}} = 1 - 2r^2 + r^2 = 1 - r^2$$

This result, which we return to in Regression 3, is extremely important. It demonstrates that we can interpret r^2 as the proportion of the variability of the dependent variable (represented by SS_{yy}) that is explained by the predictor variable (SS_{RES} represents the unexplained residual variability).

2.7.2 Relationship between *r* and the test statistic relating to the null hypothesis that the slope of the relationship is zero

There is a relationship between the Pearson correlation coefficient and the *t*-statistic used to test H_0: $\beta = 0$ introduced in 2.6.1. The relationship is

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad (16)$$

Proof:

From (15) and $r^2 = \frac{S_{xy}^2}{SS_{xx}SS_{yy}}$:

$$\frac{r^2}{1-r^2} = \frac{S_{xy}^2}{SS_{xx}SS_{yy}} \Big/ \frac{SS_{RES}}{SS_{yy}}$$

$$= \frac{S_{xy}^2}{SS_{xx}SS_{RES}}$$

$$= \frac{S_{xy}^2}{SS_{xx}(n-2)\hat{\sigma}^2}$$

Substituting in from (8) $\hat{\sigma}^2 = V(\hat{\beta})SS_{xx}$

$$= \frac{S_{xy}^2}{SS_{xx}^2(n-2)V(\hat{\beta})}$$

$$= \frac{\hat{\beta}^2}{(n-2)V(\hat{\beta})}$$

$$= \frac{t^2}{(n-2)}$$

This justifies the use of *t* as a statistic that can be used to test H_0:ρ =0 as stated in Analytical Techniques 4. Further this result shows that the test of association between two variables can be carried out either through calculation and testing of the Pearson correlation coefficient or through the use of a simple linear regression model. As is illustrated by the example in 2.4.2 the statistical significance of the slope is unaffected by which variable is chosen as the predictor variable.

Regression 3: Introduction to Analysis of Variance

3.1 Objectives

By the end of this session students will be able to:

- State the purpose of Analysis of Variance (ANOVA).
- Perform and interpret ANOVA for the simple linear regression model both by hand calculation and using Stata.
- Explain the relationship between the *F* and *T* statistics for the simple linear regression model.
- Perform and interpret a one-way ANOVA both by hand calculation and using Stata.

3.2 Introduction

When fitting statistical models we often wish to compare the extent to which two models **fit** the data. Sometimes we may wish to assess the fits of comparably complex models. For example we may wish to consider whether systolic or diastolic blood pressure better predicts a person's risk of suffering a stroke. In this example we are interested in comparing the fits of two models, each with a single predictor variable. In fact questions of this type are not straightforward to answer because the two models are not **nested**.

Statisticians refer to models being **nested** when one of the models (the **simpler** model) contains a subset of the predictor variables in the other one (the **complex** model) and no additional predictor variables. The comparison of nested models is much more straightforward from a statistical point of view than that of non-nested models and accordingly in the Foundations module we only consider the comparison of nested models.

The key to the statistical comparison of nested **linear** regression models is **Analysis of Variance (ANOVA)**. In term 2 you will be introduced to a larger class of models termed **generalised linear models**. The fits of nested generalised linear models are compared using **likelihood ratio tests**. The relationship between the test statistics that are derived from Analysis of Variance and likelihood ratio test statistics will be discussed in the generalised linear models module.

The principle of ANOVA is that if the complex model better describes the data than the simpler model then we would expect that a 'reasonably large' amount of the residual variation that is unexplained by the simpler model will be explained by the complex one. Moe formally we would expect the residual sum of squares for the complex model to be 'substantially less' than that for the simpler model. ANOVA allows us to give formal statistical meaning to concepts like 'reasonably large' and 'substantially less'.

In the first part of this session ANOVA will be introduced in the simple context of a comparison of the fit of a simple linear regression model with a simpler model that postulates no association between the dependent and predictor variables (the **null** model). Developing ANOVA in this simple context illustrates the concepts that are used in more complex settings later.
Of course another way to compare the fits of the simple linear regression and the null models is to conduct a Wald test the null hypothesis that the slope parameter in the linear regression

model is zero. In section 3.3.6 we demonstrate that the two approaches are mathematically equivalent.

3.3 ANOVA for the simple linear regression model

As explained above ANOVA provides a method of comparing the fit of two or more mean functions for the same data. For the Gambian growth data (example 1) we compare the fits of:

i) a **null model** that assumes no relation between weight and age and
ii) an **alternative model** that assumes a linear relation between weight and age.

These two models are shown in Figure 1.

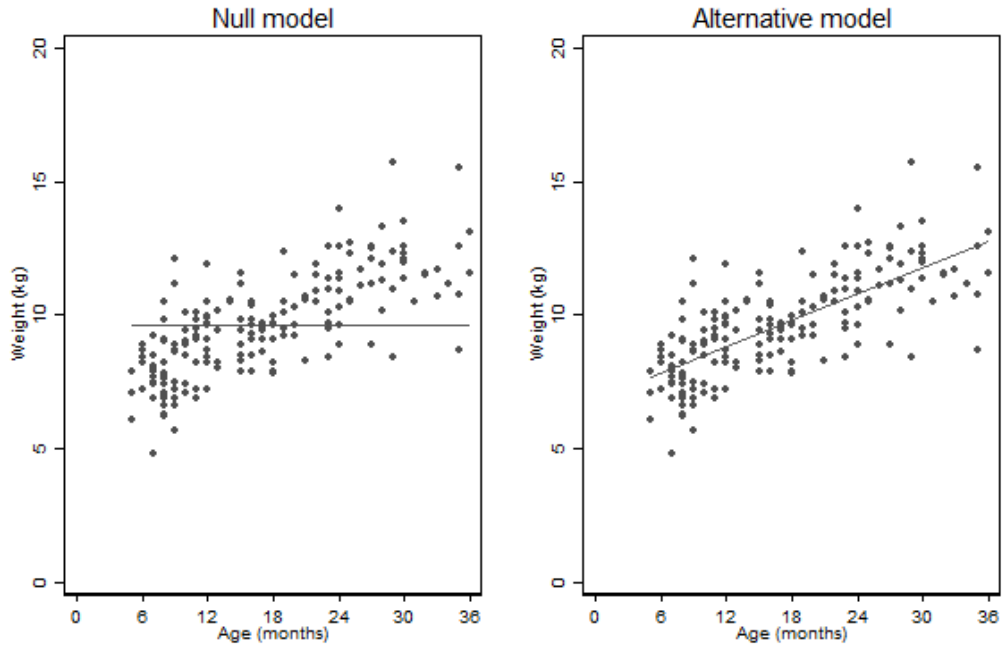


Figure 1: "Null" and "Alternative Models" for the data from the Gambian cross-sectional survey.

3.3.1 Parameter estimates for the two models

For both the null and alternative models estimates are found by minimising the residual sum of squares.

$$SS_{RES} = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

For the alternative model the mean function is assumed to follow a linear relationship of the form $\mathrm{E}(Y \mid X = x) = \alpha + \beta x$ and parameter estimates and the residual sum of squares are as defined in Regression 1 (equations (5)-(7)).

The null model specifies the mean function to be $E(Y|X = x) = \alpha$. To fit this model we find the best line parallel to the *X*-axis by minimizing:

$$SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{\alpha})^2$$

Since $\hat{\alpha} = \bar{y} - \hat{\beta}$ from regression 1 equation (6), this gives $\hat{\alpha} = \bar{y}$

It further follows that, for this model:

$$SS_{RES} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_{yy}$$

i.e. the residual sum of squares for the null model is the corrected sum of squares of the *Y*'s.

3.3.2 Partitioning the residual sum of squares from the null model

The principle behind Analysis of Variance is that the Residual Sum of Squares from the simpler model (here the null model), denoted by SS_{RES_NULL}, can be **partitioned** into two parts. These are:

i) the sum of squares explained by the complex model (here the simple linear regression model), SS_{REG} and
ii) the residual sum of squares from the complex model, SS_{RES_ALT}.

The key result is:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

$$SS_{RES\_NULL}(SS_{yy}) = \quad SS_{REG} \quad + \quad SS_{RES\_ALT} \tag{1}$$

Proof:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} ((\hat{y_i} - \bar{y}) + (y_i - \hat{y_i}))^2$$

$$= \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + 2\sum_{i=1}^{n} (\hat{y_i} - \bar{y})(y_i - \hat{y_i})$$

Hence it is required to prove that $\sum_{i=1}^{n} (\hat{y_i} - \bar{y})(y_i - \hat{y_i}) = 0$

Since $\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ (regression 2, equation (4)) it follows that

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}\left(\bar{y} + \hat{\beta}(x_i - \bar{x}) - \bar{y}\right)(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))$$

$$= \sum_{i=1}^{n}\hat{\beta}(x_i - \bar{x})((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))$$

$$= \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{S_{xy}}{SS_{xx}}S_{xy} - \left(\frac{S_{xy}}{SS_{xx}}\right)^2 SS_{xx}$$

$= 0$ as required

3.3.3 R^2 - The Coefficient of Determination

In equation (1) the variation explained by the simple linear regression model is SS_{REG} and that not explained is SS_{RES_ALT} (usually abbreviated by SS_{RES}). The proportion of variation which is explained by the model is denoted by $\boldsymbol{R^2}$ and termed the **Coefficient of Determination**. *i.e.*

$$R^2 = \frac{SS_{REG}}{SS_{yy}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2)$$

It is shown in Regression 2 (2.7.1) that the coefficient of determination is equal to the square of the Pearson correlation coefficient in the simple linear regression model. An alternative derivation, following from equation (2) is:

Proof:

Since $(\hat{y}_i - \bar{y}) = \hat{\beta}(x_i - \bar{x})$
and $\hat{\beta} = S_{xy}/SS_{xx}$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\hat{\beta}^2 SS_{xx}}{SS_{yy}} = \frac{S_{xy}^2}{SS_{xx}SS_{yy}} = r_{xy}^2 \qquad (3)$$

Example: For the regression model relating weight to age in the Gambian cross-sectional study the coefficient of determination (R^2) is 0.54. The interpretation is that (a linear relationship with) age explains 54% of the variability in children's weight in this data. The coefficient of determination is presented in the output from Stata's **regression** command as shown below.

```
. regress wt age

      Source |       SS       df       MS      Number of obs =     190
-------------+------------------------------   F(  1,   188) =  221.39
       Model |  359.063204     1  359.063204   Prob > F      =  0.0000
    Residual |  304.906554   188  1.62184337   R-squared     =  0.5408
-------------+------------------------------   Adj R-squared =  0.5383
       Total |  663.969759   189  3.51306751   Root MSE      =  1.2735
```

3.3.4 The ANOVA table

The result of partitioning the sum of squares as described in the previous sections is usually displayed in an Analysis of Variance table. For a simple linear regression model this is as follows.

Source of variation due to	Sum of Squares	Degrees of Freedom	Mean Sum of Squares
Regression (model)	SS_{REG}	1	$MS_{REG} = SS_{REG} / 1$
Residual	SS_{RES} (SS_{RES_ALT})	n-2	$MS_{RES} = SS_{RES}/(n\text{-}2)$
Total	SS_{yy} (SS_{RES_NULL})	n-1	$SS_{yy}/(n\text{-}1)$

Table 1: Analysis of Variance table for a simple linear regression model

The final column in this table shows the **mean sums of squares**. In each case these are the sum of squares divided by the appropriate degrees of freedom. The residual mean sum of squares MS_{RES} is the unbiased estimator of the residual variance ($\hat{\sigma}^2$) under the alternative model (here the simple linear regression model). The total mean sum of squares ($S_{yy}/(n\text{-}1)$) is estimator of the residual variance under the null model.

The analysis of variance table is presented in the output from Stata's **regression** command as shown above for the Gambian growth data.

3.3.5 Hypothesis testing using ANOVA

The test that accompanies an ANOVA table is the *F* test. In essence this compares the size of the sum of squares explained by the complex model (SS_{REG}) and the residual sum of squares from the complex model (SS_{RES}) to see whether the size of SS_{REG} is compatible with the play of chance under the null hypothesis that the simpler model is correct.

3.3.5.1 Distributions of sums of squares

To construct the *F* test we need to know the distributions of SS_{REG} and SS_{RES} under the null hypothesis ($H_0 : \beta = 0$) and alternative hypotheses ($H_1: \beta \neq 0$). These are as follows:

i) Under both H_0 and H_1 SS_{RES} follows a multiple of a χ^2 distribution with *n*-2 degrees of freedom.

$$\frac{SS_{RES}}{\sigma^2} \sim \chi^2_{n-2} \tag{4}$$

ii) Under H_0 SS_{REG} follows a multiple of a χ^2 distribution with 1 degree of freedom that is **independent of SS_{RES}**.

$$\frac{SS_{REG}}{\sigma^2} \sim \chi^2_1 \tag{5}$$

iii) Under H_1 SS_{REG} follows a non-central χ^2 distribution that is **independent of SS_{RES}**.

$$SS_{REG} = \beta^2 SS_{xx} + U \quad \text{where } \frac{U}{\sigma^2} \sim \chi^2_1 \tag{6}$$

(Note that for a simple linear regression model $MS_{REG} = SS_{REG}/1 = SS_{REG}$)

3.3.5.2 The *F* test in simple regression

We can see from the above that under the null hypothesis both SS_{RES} and SS_{REG} follow distributions that are multiples of independent χ^2 distributions. As was discussed in Analytical techniques 5, if two random variables follow independent χ^2 distributions then their ratio scaled by the respective degrees of freedom of these distributions follows an F-distribution. In other words if $A \sim \chi^2_a$ and $B \sim \chi^2_b$:

$$\frac{A/a}{B/b} \sim F_{a,b}$$

So, it follows that under the null hypothesis (H_0) the ratio of MS_{REG} to MS_{RES} will follow an *F* distribution. Specifically

$$F = \frac{SS_{REG}/1}{SS_{RES}/(n-2)} = \frac{MS_{REG}}{MS_{RES}} \sim F_{1,(n-2)} \tag{7}$$

Under the alternative hypothesis ($H_1: \beta \neq 0$) the expectation of SS_{REG} is $\sigma^2 + \beta^2 SS_{xx}$ and hence the expectation of *F* will be greater than that of $F_{1,(n-2)}$, which is approximately 1 (the median of the distribution of $F_{1,(n-2)}$ is 1, but the mean is actually 1 + (2/(*n* - 4))). Hence a comparison of *F* with the $F_{1,(n-2)}$ distribution yields a *p*-value for testing the null hypothesis.

It is important to note that it is only probabilities in the upper tail of the appropriate *F* distribution that are relevant here because departures from the null hypothesis always result in the expectation of *F* being greater than 1. This can be a source of confusion in describing the test, but should not be. The test is a **two-sided** one (the alternative hypothesis is H_1: $\beta \neq 0$, not H_1: $\beta > 0$ or H_1: $\beta < 0$), but the *p*-value is calculated from a one-sided comparison of the test statistic with the $F_{1,(n-2)}$ distribution.

The output below shows calculation of the *F* statistic and its associated p-value in Stata for the Gambian cross-sectional survey.

```
. regress wt age

      Source |       SS       df       MS       Number of obs =     190
-------------+------------------------------    F(  1,   188) =  221.39
       Model |  359.063204     1  359.063204    Prob > F      =  0.0000
    Residual |  304.906554   188  1.62184337    R-squared     =  0.5408
-------------+------------------------------    Adj R-squared =  0.5383
       Total |  663.969759   189  3.51306751    Root MSE      =  1.2735

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1653314   .0111115    14.88   0.000     .1434121    .1872507
       _cons |   6.837584   .2100701    32.55   0.000     6.423187    7.251982
------------------------------------------------------------------------------
```

Exercise: Calculate the *F* statistic from the appropriate entries in the ANOVA table for the Gambian cross-sectional data.

3.3.6 Equivalence of *F* and *t* tests in simple linear regression

As might be anticipated, for simple linear regression, the *F* and *t* tests lead to identical *p*-values and hence can be considered to be the same statistical test. This can be shown as follows:

Proof:

$$F = \frac{SS_{REG}/1}{SS_{RES}/(n-2)} = \frac{SS_{REG}}{(SS_{yy} - SS_{REG})/(n-2)}$$

Since $r^2 = SS_{REG}/SS_{YY}$:

$$F = (n-2)\left(\frac{SS_{yy}r^2}{SS_{yy} - SS_{yy}r^2}\right) = (n-2)\left(\frac{r^2}{1-r^2}\right) = t^2$$

(See Regression 2 section 2.7.2 for a reminder of the relationship between *r* and *t*.)

The equivalence of the two tests follows from the fact that $F_{1,(n-2)} = t^2_{(n-2)}$

Exercise: Demonstrate the relationship between the *F* and *t* statistics for the Gambian cross-sectional data.

3.4 ANOVA for models with categorical predictor variables

If Analysis of Variance only provided an alternative approach to carrying out a test of association in the simple linear regression model then it would be of limited interest. However ANOVA has much greater utility than this, both for linear regression models involving more than one continuous predictor (see Regression 4), models with one or more categorical predictor variables, and models involving both categorical and continuous predictors.

As we have seen in Regression 1, models with a binary predictor variable can be treated as linear regression models by using "dummy" variables. The same approach can be used for any linear regression model that includes binary or categorical predictor variables. However such models are traditionally described with different terminology and somewhat different statistical notation from that used to describe linear regression models. In this section we show the links between these two nomenclatures.

3.4.1 A simple case: one binary predictor variable

It has been demonstrated in Regression 1 that, through careful choice of the coding of the two values that a binary predictor variable can take (0 and 1), linear regression can be used to investigate the relationship between a binary predictor variable and a continuous outcome, giving results that are equivalent to those from a t-test.

We return to the example used in Regression 1. Eleven newborns were randomised to two groups, an eight week active exercise group and an eight week control group and ages at time of first walking compared between the groups. Figure 2 shows the data.

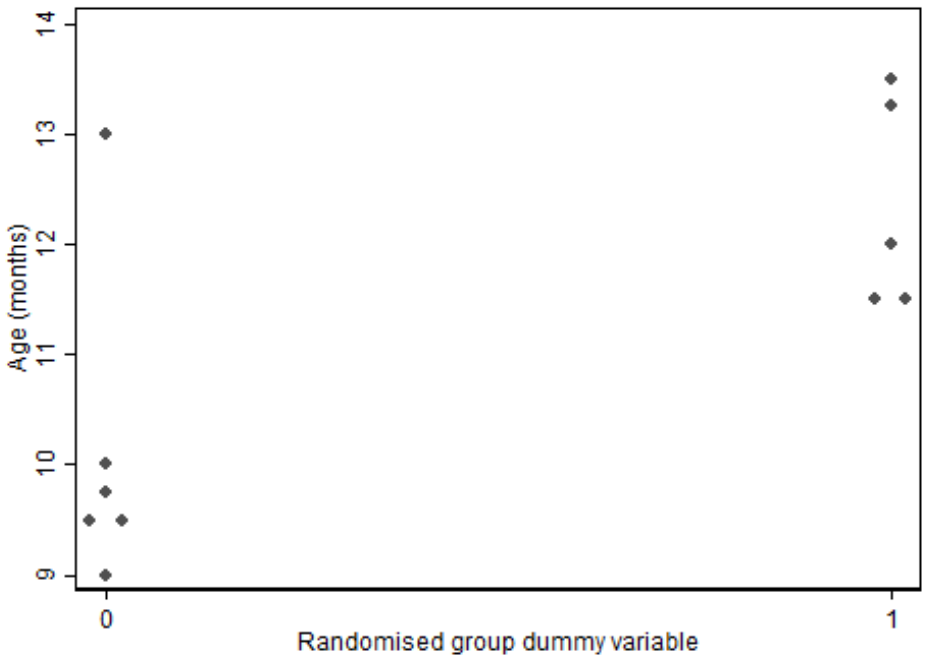


Figure 2: Exercise regime trial: age at walking by randomisation group. Points representing equal walking ages have been 'jittered' (moved a small distance to the left or right) to aid clarity. The 'dummy' variable takes the value 0 in the active exercise group and 1 in the control group.

The following output shows the results obtained using the **regress** command in Stata.

```
. regress age group

      Source |       SS       df       MS        Number of obs =      11
-------------+------------------------------     F(  1,     9) =    8.58
       Model |  13.5017045     1  13.5017045     Prob > F      =  0.0168
    Residual |    14.16875     9  1.57430556     R-squared     =  0.4879
-------------+------------------------------     Adj R-squared =  0.4311
       Total |  27.6704545    10  2.76704545     Root MSE      =  1.2547

------------------------------------------------------------------------
      age |   Coef.  Std. Err.   t    P>|t|  [95% Conf. Interval]
----------+-------------------------------------------------------------
    group |    2.225  .7597667   2.93  0.017  .5062884    3.943712
    _cons |  10.125  .5122346  19.77  0.000  8.966245    11.28376
------------------------------------------------------------------------
```

As explained in Regression 1 the slope parameter (**group**) is here interpretable as the difference in the mean ages between the two groups. Note that the *p*-value from a test of the null hypothesis that the slope is zero is the same (0.017) as that from the Analysis of Variance (although Stata gives an extra decimal place to the *p*-value from the *F* test).

3.4.1.1 Two formulations of the model

There are two ways of algebraically specifying the model relating a single binary predictor to a continuous outcome. The representation that tallies with the simple linear regression model used for a single continuous predictor is as follows.

Let y_i and x_i be the values of the outcome ('age at walking' for the exercise trial example) and predictor variables ('group' for the example) for the *i*th subject (i=1,..,n) (n = 11 for the example). Then **regression representation** of the model is as follows.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \text{where } \varepsilon_i \sim NID(0, \sigma^2) \qquad (8)$$

Where,
$x_i = 0$ if the *i*th newborn is in the 'active' group
$x_i = 1$ if the *i*th newborn is in the 'control' group

With this (regression) representation the null hypothesis is H_0: $\beta = 0$ and the alternative hypothesis is H_1: $\beta \neq 0$

The alternative representation, sometimes termed the **ANOVA representation**, is as follows:

Let y_{ki} be the value of the outcome for the *i*th subject in the *k*th group (i=1,.., n_k and k = 1,..,K). For the exercise trial data there are 2 groups (K= 2), with 6 infants in the 'active' group (n_1 = 6) and 5 infants in the 'control' group (n_2 = 5). The model is as follows.

$$y_{ki} = \mu_k + \varepsilon_{ki}, \text{where } \varepsilon_{ki} \sim NID(0, \sigma^2) \qquad (9)$$

Here μ_k is the mean of the dependent variable in the *k*th group. With this representation the null hypothesis is H_0: $\mu_k = \mu$ (means in all groups are equal to a common value μ) and the alternative hypothesis is H_1: $\mu_k \neq \mu$.

Exercise: Write down the mathematical relationship between the parameters α and β in (8) and μ_1 and μ_2 in (9) i.e. $\alpha = \ldots \quad \beta = \ldots$

3.4.1.2 Sums of squares for models with categorical predictors.

For models with a single categorical predictor variable the fitted values are simply the group specific means ($\bar{y}_k$). Under the null hypothesis that the group means are all equal, the fitted values are all equal to the overall mean ($\bar{y}$). This leads to new terminology for the residual sum of squares (SS_{RES}) and the sum of squares explained by the model (SS_{REG}). Using the notation of equation (9) the residual sum of squares is given by the following.

$$SS_{RES} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ki} - \bar{y}_k)^2 \qquad (10)$$

With such models the residual sum of squares is often termed the **within group sum of squares**.

The sum of squares explained by the model is

$$SS_{REG} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\bar{y}_k - \bar{y})^2 = \sum_{k=1}^{K} n_k \, (\bar{y}_k - \bar{y})^2 \qquad (11)$$

This is often termed the **between group sum of squares**. This terminology is used in Stata's **`oneway`** command, as illustrated below for the exercise trial data.

```
. oneway age group

                    Analysis of Variance
    Source              SS       df      MS         F    Prob > F
------------------------------------------------------------------
Between groups    13.5017045     1   13.5017045   8.58   0.0168
 Within groups      14.16875     9   1.57430556
------------------------------------------------------------------
    Total         27.6704545    10   2.76704545

Bartlett's test for equal variances:  chi2(1) =   0.6302  Prob>chi2 = 0.427
```

Notice that the ANOVA table and the reported *F* statistic are identical to those resulting from the use of the **`regress`** command. For information, Bartlett's test is an alternative to the test of equal variances introduced in Analytical Techniques 5.

3.4.2 Models with a single categorical predictor variable taking three or more values

Both the **regression** and **ANOVA** approaches to the analysis of a model with a single binary predictor can be extended to handle models with a categorical predictor having three or more levels. The regression approach (equation (7)) requires a multivariable model, which will be introduced in Regression 4. The ANOVA approach (termed **one-way Analysis of Variance)**, requires only a little modification, which we will describe here.

Equations (9), (10) and (11) are general statements that hold for K groups and can therefore be used without modification. However when there are K groups the degrees of freedom for the within groups (*i.e.* residual) sum of squares are $n - K$ (because the model includes K parameters) whilst that for the between group sum of squares is $K - 1$ (because the null model contains a single parameter, the overall mean). Hence the Analysis of Variance table is as follows:

Source of variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares
Between groups	$SS_{BETWEEN}$	$K - 1$	$SS_{BETWEEN}/(K - 1)$
Within Groups	SS_{WITHIN}	$n - K$	$SS_{WITHIN}/(n-K)$
Total	SS_{yy} (SS_{RES_NULL})	n-1	$SS_{yy}/(n-1)$

Table 2: One-way Analysis of Variance table.

The F statistic is

$$F = \frac{SS_{BETWEEN}/(K-1)}{SS_{WITHIN}/(n-K)} \sim F_{(K-1),(n-K)} \text{ under } H_0 \qquad (12)$$

Care must be taken in interpreting the result of such an F test. A small p-value provides evidence that the means in the groups are not all the same. This is not the same thing as providing evidence that all of the group means are different. In particular, with more than two groups it does not tell us which of the group means differed from which other group means. For this reason if we find evidence of difference in means on an F test, we would almost always follow this up by further analysis. Such further analysis might involve pair-wise comparison of means either through analysis restricted to two groups, or through the use of **contrasts** (see session on strategies of analysis in the Generalised Linear Models module for further details).

The great advantage of a one-way analysis of variance is that it provides a single test statistic to assess the evidence that mean levels in three or more groups are the same. An alternative approach would be to use a series of pair-wise comparisons. However this introduces problems of interpretation since making multiple comparisons increases the possibility of false positive results. For example with five groups, ten pair-wise comparisons can be made. So even if the null hypothesis is true and all groups have the same mean, the probability of $p < 0.05$ for at least one of the tests would be much greater than 5%. The safest approach with such data is to first carry out the F-test and then only to investigate pair-wise comparisons when the first F-test has a small p-value. Such a strategy may not be the best one if, a-priori, the scientific question of interest is that (for example) one group has very different mean levels from each of the others, with the other mean levels all anticipated to be roughly equal. This topic will be returned to in sessions on strategies of analysis in later modules.

Example: Vital Capacity in workers exposed to cadmium fumes

A one-way analysis of variance is illustrated using data from a study investigating whether vital capacity (the maximum volume of air that a person can exhale after maximum inhalation) is reduced in workers at a factory who are exposed to cadmium fumes. Figure 3 shows vital capacity (litres) in three groups of workers: i) those not exposed to cadmium fumes, ii) those exposed for less than 10 years and iii) those exposed for more than 10 years.

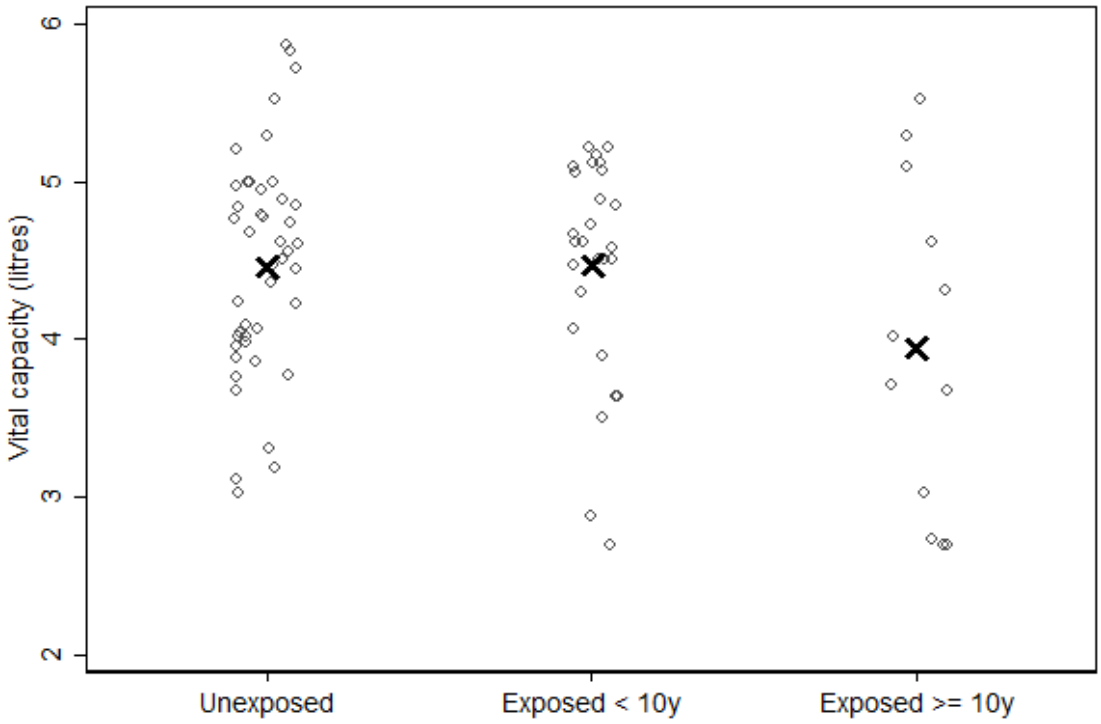


Figure 3. Observational study of vital capacity in workers exposed to cadmium fumes and unexposed controls. Group-specific mean levels shown as crosses.

The following Stata output shows the results of fitting a one-way Analysis of Variance Model to this data.

```
 oneway vitcap group

                    Analysis of Variance
    Source              SS       df      MS         F    Prob > F
------------------------------------------------------------------
Between groups    2.74733751     2   1.37366875   2.48     0.0902
 Within groups    44.8936168    81    .554242182
------------------------------------------------------------------
    Total         47.6409543    83    .573987401
```

Here from the *F* test we have p=0.09, which provides only weak evidence against the null hypothesis that the means in the three groups are equal.

Regression 4: Introduction to Multivariable Models

4.1 Objectives

By the end of this session students will be able to:

- Explain the motivation for multivariable linear regression models.
- Fit and interpret linear regression models with two predictor variables using Stata.
- Fit and interpret ANOVA and ANCOVA models.
- Explain the reasons for changes in parameter estimates when new predictor variables are added to a regression model.

4.2 Introduction

The simple linear regression model describes the relationship between a **continuous** dependent variable and a single predictor variable. Often we wish to extend this model to provide a statistical framework to describe the way in which two or more predictor variables relate to a dependent variable. For example we may wish to fit a model that relates various components of a person's life-style (age, gender, exercise, diet, etc.) to their systolic blood pressure. There can be a number of different motivations for such an analysis. In some situations we may be interested in the joint dependency on all the predictor variables. Alternatively we may be particularly interested in the relationship between the dependent variable and a particular predictor variable (for example, exercise (staying with the above setting)) but need to **control** (or **adjust**) for the **confounding** effects of other determinants (age, gender, diet, *etc.*). Alternatively we may not be particularly interested in the form of the relationship, but merely wish to be able to **predict** the outcome variable from a number of other variables (for example if we wanted to develop a tool to predict someone's blood pressure without having to directly measure it).

In general it is always a good idea to consider the purpose of any statistical analysis before fitting a model; to make sure that the model that is used, and the inferences made from it, correspond to the aims of the study. For example if we are interested in investigating whether a 'newly discovered' factor is related to blood pressure, after adjusting for age, gender and other known determinants, then the analysis should focus on inferences concerning the new factor. Here *p*-values relating to the null hypothesis that age is not related to blood pressure (for example) would be of little interest.

Multivariable linear regression (often termed **multiple linear regression**) is the most widely used technique for investigating the relationship between a single continuous dependent variable and two or more predictor variables and is the subject of this session. In the first half of the session the multiple linear regression model will be presented. However the real challenges for statistical and non-statistical researchers alike lies in the interpretation of results from linear regression models with many predictor variables. The last part of the session will focus on this, particularly on the important concept of **confounding**.

4.3 Linear regression models with two predictors

4.3.1 Formulation and interpretation of the model

Suppose we wish to relate a dependent variable (Y) to two predictor variables (X_1 and X_2). For example for the Gambian cross-sectional survey we might wish to relate a child's weight (Y) to their age (X_1) and their length (X_2). The linear regression model here is as follows.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \text{where } \varepsilon_i \sim NID(0, \sigma^2) \qquad (1)$$

Where,

y_i = value of the dependent variable (e.g. weight) for the *i*th observation (e.g. *i*th child participating in the study)

x_{1i} = value of the first predictor variable (e.g. age) for the *i*th observation

x_{2i} = value of the second predictor variable (e.g. length) for the *i*th observation

The parameters in the model are as interpreted as follows:

α is the intercept. It is the expectation of Y when $\boldsymbol{X_1}$ **and** $\boldsymbol{X_2}$ **are both zero**.

β_1 is the increase in the expectation of Y for a **1 unit increase in** $\boldsymbol{X_1}$ **with** $\boldsymbol{X_2}$ **held constant**.

β_2 is the increase in the expectation of Y for a **1 unit increase in** $\boldsymbol{X_2}$ **with** $\boldsymbol{X_1}$ **held constant**.

β_1 and β_2 are called **partial** regression coefficients. They measure the effect of one predictor variable **controlled** (or **adjusted**) for the other.

In order to conceptualise this model it can be helpful to visualise the data in three dimensional space, with axes defined by X_1, X_2 and Y (with Y as the vertical axis). If we visualise the data in this way then fitting model (1) identifies the **plane** of best fit *i.e.* the choice of plane that minimises the sum of squared vertical deviations from itself.

This model can also be expressed using matrix algebra.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \text{where } \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}\sigma^2) \qquad (2)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & . & . \\ 1 & . & . \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

In this formulation $\boldsymbol{X}$ is a matrix and each of $\boldsymbol{Y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are vectors. The residuals are assumed to follow a **multivariate normal distribution** with variance-covariance matrix equal to σ^2 multiplied by the identity matrix ($\boldsymbol{I}$). This is equivalent to assuming that residuals are normally and independently distributed with constant variance σ^2.

4.3.2 Least Squares Estimation

As with simple linear regression parameters estimates are obtained by minimising the residual sum of squares. For a model with two predictors this is

$$SS_{RES} = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2 \tag{3}$$

Just as with simple linear regression this leads to closed form solutions for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$. Each of these estimators is unbiased and there are closed form solutions for their variances and covariances. Furthermore it can be shown that the following is an unbiased estimator for σ^2.

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^{\,2}}{(n-3)} = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2 / (n-3) \tag{4}$$

These formulae are given (using matrix formulation) in Regression 5, which also covers significance tests and construction of confidence intervals. In practice, statistical software can be used to take care of all the calculations. In the rest of this session we will consider the interpretation of the results from multivariable models, which is where the primary challenge for the statistician lies.

4.3.3 Example: REPAIR clinical trial

The renal protection against ischaemia-reperfusion in transplantation (REPAIR) randomised controlled clinical trial was carried out in 400 living-donor renal transplant patients (http://repair.lshtm.ac.uk/). Here Stata is used to relate recipient's glomerular filtration rate (GFR) at 12 months after transplant (`gfr_iohexol`) to donor's age in years (`d_age_b`) and donors weight in kg (`d_weight_b`), both measured at the start of the study.

Model 1:

```
. regress gfr_iohexol d_age_b

      Source |       SS           df       MS      Number of obs   =       309
-------------+----------------------------------   F(1, 307)       =     58.72
       Model |   13007.521         1   13007.521   Prob > F        =    0.0000
    Residual |  68009.2509       307  221.528505   R-squared       =    0.1606
-------------+----------------------------------   Adj R-squared   =    0.1578
       Total |  81016.7719       308  263.041467   Root MSE        =    14.884

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     d_age_b |  -.5170756   .0674795    -7.66   0.000    -.6498563   -.3842948
       _cons |   84.76726   3.461534    24.49   0.000     77.95593    91.57859
------------------------------------------------------------------------------
```

```
. regress gfr_iohexol d_weight_b
```

Model 2:

```
      Source |       SS           df       MS      Number of obs   =       309
-------------+----------------------------------   F(1, 307)       =      8.53
       Model |  2189.97013         1  2189.97013   Prob > F        =    0.0038
    Residual |  78826.8017       307  256.764827   R-squared       =    0.0270
-------------+----------------------------------   Adj R-squared   =    0.0239
       Total |  81016.7719       308  263.041467   Root MSE        =    16.024

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  d_weight_b |    .188598   .0645782     2.92   0.004     .0615261    .3156698
       _cons |   44.61107   5.026812     8.87   0.000     34.71971    54.50244
------------------------------------------------------------------------------
```

```
. regress gfr_iohexol d_age_b d_weight_b
```

Model 3:

```
      Source |       SS           df       MS      Number of obs   =       309
-------------+----------------------------------   F(2, 306)       =     34.46
       Model |  14894.5754         2  7447.28769   Prob > F        =    0.0000
    Residual |  66122.1965       306  216.085609   R-squared       =    0.1838
-------------+----------------------------------   Adj R-squared   =    0.1785
       Total |  81016.7719       308  263.041467   Root MSE        =      14.7

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     d_age_b |  -.5112434   .0666745    -7.67   0.000     -.642442   -.3800448
  d_weight_b |   .1751461   .0592681     2.96   0.003     .0585214    .2917708
       _cons |   71.06971    5.75955    12.34   0.000     59.73637    82.40304
------------------------------------------------------------------------------
```

Parameters estimates, standard errors and the estimated residual variance all differ between the models.

$\hat{\alpha}$ = 71.1 in model 3 is the estimated intercept. The interpretation is that the estimated mean GFR of a transplant recipient whose donor was aged 0 years and had weight 0kg is 71.1 mL/min per 1.73 m^2. This (clearly ridiculous) estimate is an **extrapolation** based on the associations for donors of older ages and non-zero weights and an assumption of linearity. Even though the this intercept has no practical meaning in of itself, this does not mean that the model is unreasonable for recipients whose donors were of ages and weights that are typical of the sample included in the REPAIR clinical trial.

$\hat{\beta}_1$ = -0.511 in model 3 is the estimated partial regression coefficient for donor age. The interpretation is that the estimated mean increase in a recipient's GFR per one year increase in the donor's age **amongst recipients whose donor was of the same weight** is -0.511 mL/min per 1.73 m^2. Based on the p-value ($p<0.001$) we have good evidence that there is a negative association between donor age and recipient GFR. For a 10 year increase in donor age the expected decrease in recipients GFR is around 5 mL/min per 1.73 m^2. Notice that the regression coefficient for donor age is slightly different between Model 1 (-0.517) and Model 3 (-0.511). The reason for the difference is that the coefficient from Model 1 is the estimated mean increase in GFR per year increase in donor age **ignoring donor weight**. While the multiple regression model (Model 3) addresses the question of whether recipients with older donors have different GFR, on average, than those with younger donors of the same weight.

$\hat{\beta}_1$ = 0.175 in model 3 is the estimated partial regression coefficient for donor weight. The interpretation is that the estimated mean increase in a recipient's GFR per one kg increase in the donor's weight **amongst recipients whose donor was of the same age** is 0.175 mL/min per 1.73 m^2. We have good evidence (p=0.003) that higher donor weight is associated with higher recipient GFR at 12 months after transplant. Notice that the again the regression coefficient for donor weight is different in Model 3 than in Model 2. The reason for the difference is that the coefficient from Model 2 is the estimated mean increase in GFR per kg increase in donor weight **ignoring donor age**.

Often researchers attempt to make causal inferences from regression models such as these. However, this can be far from straightforward. Based on the results presented above it would be reasonable to say that 'in recipients of live-donor kidney transplant, GFR at 12 months after transplant decreased with donor age and increased with donor weight'. However, it is much less reasonable to claim that the lower GFR in recipients with older donors is caused by the increased donor age. There are many other possibilities, such as the fact that if the donor was older then the recipient may also be older and this might influence their GFR. Causal statements require more than just the results of a statistical model to make them plausible. We return to this in Section 4.7.

We have introduced multiple regression models with an example where both predictors are continuous. Before focussing on changes in estimated regression coefficients when new predictor variables are added we consider two types of multiple regression model that include categorical predictors. These models are commonly called ANOVA and ANCOVA models.

4.4 Categorical predictor variables in linear regression models

It has already been demonstrated (Regression 1) that binary predictor variables can be included in linear regression models through the creation of a 'dummy' variable taking the value 0 and 1 in the two groups. With a categorical predictor variable taking three or more values the extension of this is to create more than one dummy variable. We illustrate this approach for a categorical variable (*X*) taking three possible values (1, 2, 3). The model is as follows

$$y_i = \alpha + \beta_1 u_{1i} + \beta_2 u_{2i} + \varepsilon_i, where\ \varepsilon_i \sim NID(0, \sigma^2) \qquad (5)$$

Where,

$$u_{1i} = \begin{cases} 1\ if\ x_i = 2 \\ 0\ if\ x_i \neq 2 \end{cases}$$

$$u_{2i} = \begin{cases} 1\ if\ x_i = 3 \\ 0\ if\ x_i \neq 3 \end{cases}$$

The reason why the model is formulated in this way becomes apparent if we note that equation (5) can be written as follows:

$y_i = \alpha + \varepsilon_i$ if $x_i = 1$

$y_i = \alpha + \beta_1 + \varepsilon_i$ if $x_i = 2$

$y_i = \alpha + \beta_2 + \varepsilon_i$ if $x_i = 3$

This makes explicit the interpretation of the parameters in the model.

α is the expectation of Y when $\boldsymbol{X}$**=1**.

$\alpha + \beta_1$ is the expectation of Y when $\boldsymbol{X}$**=2**. Hence β_1 is the **differenc**e in the expectation of Y between groups defined by $\boldsymbol{X}$**=2** and $\boldsymbol{X}$**=1**.

$\alpha + \beta_2$ is the expectation of Y when $\boldsymbol{X}$**=3**. Hence β_2 is the **differenc**e in the expectation of Y between groups defined by $\boldsymbol{X}$**=3** and $\boldsymbol{X}$**=1**.

In this parameterisation of the model the group defined by X=1 is often referred to as the **baseline group**. There is no statistical reason why one group rather than another should be chosen as the baseline group and it can sometimes be desirable to **re-parameterise** a model of this type to estimate parameters representing differences in mean levels from a particular baseline group.

When continuous data in three groups is displayed it is natural to display it in two dimensions with group along the X-axis and the continuous outcome on the Y-axis. However, to conceptualise the model we can imagine the data in three dimensions with the outcome on the vertical axis and all the points at either (0, 0) (group 1), (1, 0) (group 2) or (0, 1) (group 3) on the two perpendicular horizontal axes. The **plane** of best fit will then be the plane defined by the three group-specific mean values.

In order to fit models of this type in statistical software dummy variables need to be created. Stata can do this automatically by prefixing the names of variables by **`i.`** to instruct Stata to create dummy variables so each value of the variable is treated as a group in the model. This approach can be used with many commands in Stata, including **`regress`**. To illustrate its use we return to the Vital Capacity data introduced in Regression 3.

```
. regress vitcap i.group

    Source |       SS     df       MS              Number of obs =      84
-----------+------------------------------          F(  2,    81) =    2.48
     Model | 2.74733751   2  1.37366875            Prob > F      =  0.0902
  Residual | 44.8936168  81  .554242182            R-squared     =  0.0577
-----------+------------------------------          Adj R-squared =  0.0344
     Total | 47.6409543  83  .573987401            Root MSE      =  .74447

------------------------------------------------------------------------------
   vitcap |      Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
    group |
       2  |   .0097403  .1799744     0.05   0.957    -.3483523     .3678329
       3  |  -.5128788  .2424526    -2.12   0.037    -.9952834    -.0304741
          |
   _cons  |   4.462045  .1122337    39.76   0.000     4.238735     4.685355
------------------------------------------------------------------------------
```

As with all models with two predictor variables the global F-test here tests the null hypothesis H_0: $\beta_1 = \beta_2 = 0$. This is equivalent to specifying that the means in the three groups are all the same. The alternative hypothesis (H_1: $\beta_1 \neq 0$ and/or $\beta_2 \neq 0$) specifies that at least two of the group specific means are different. These null and alternative hypotheses are exactly the same as in the one-way ANOVA specification of this model (see Regression 3) and hence the global F-test here is equivalent to that introduced in Regression 3.

This is illustrated for the vital capacity data through comparison of the above output with that resulting from use of the **oneway** command in Stata, which carries out ANOVA.

```
. oneway vitcap group
                    Analysis of Variance
    Source          SS       df      MS          F     Prob > F
------------------------------------------------------------------------
Between groups   2.74733751     2   1.37366875    2.48     0.0902
 Within groups   44.8936168    81    .554242182
------------------------------------------------------------------------
    Total        47.6409543    83    .573987401
```

4.5 The Analysis of Covariance (ANCOVA) Model

We can fit linear regression models with both categorical and continuous explanatory variables. The simplest such model is the Analysis of Covariance (ANCOVA) model used to relate a continuous dependent variable (Y) to one continuous (X_1) and one binary (X_2, taking the values 1 and 2) predictor variable. The model is

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 u_{2i} + \varepsilon_i, \text{where } \varepsilon_i \sim NID(0, \sigma^2) \qquad (6)$$

Where,

y_{1i} = value of the continuous dependent variable (Y) for the ith observation

x_{1i} = value of the continuous predictor variable (X_1) for the ith observation

$$u_i = \begin{cases} 1 \; if \; x_{2i} = 2 \\ 0 \; if \; x_{2i} = 1 \end{cases}$$

The parameters in the model are as follows:

α is the intercept. It is the expectation of Y when **X_1 is zero and X_2 is 1 (u = 0).**

β_1 is the increase in the expectation of Y for a **1 unit increase in X_1 with X_2 held constant**.

β_2 is the **differenc**e in the expectation between groups defined by **X_2=2** and **X_2=1 with X_1 held constant**.

To interpret this model it is helpful to realise that equation (6) can be written as follows:

$y_i = \alpha + \beta_1 x_{1i} + \varepsilon_i$ if $x_{2i} = 1$

$y_i = \alpha + \beta_2 + \beta_1 x_{1i} + \varepsilon_i$ if $x_{2i} = 2$

In a two dimensional plot of Y against X_1 the fitted values appear as **parallel lines** with β_2 the distance between two lines defined by X_2 =1 and X_2 =2, and β_1 the common slope.

This model is illustrated by relating weight (kg) to age (months) and gender (male=1, female=2) in the Gambian cross-sectional data. Stata's output is follows. Figure 1 shows the data and fitted values from the model.

```
. regress wt age i.sex

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .1639984   .0109195    15.02   0.000     .1424572    .1855395
   2.sex |   -.518854   .1830531    -2.83   0.005    -.8799686   -.1577394
   _cons |   7.152414   .2342542    30.53   0.000     6.690293    7.614534
------------------------------------------------------------------------------
```

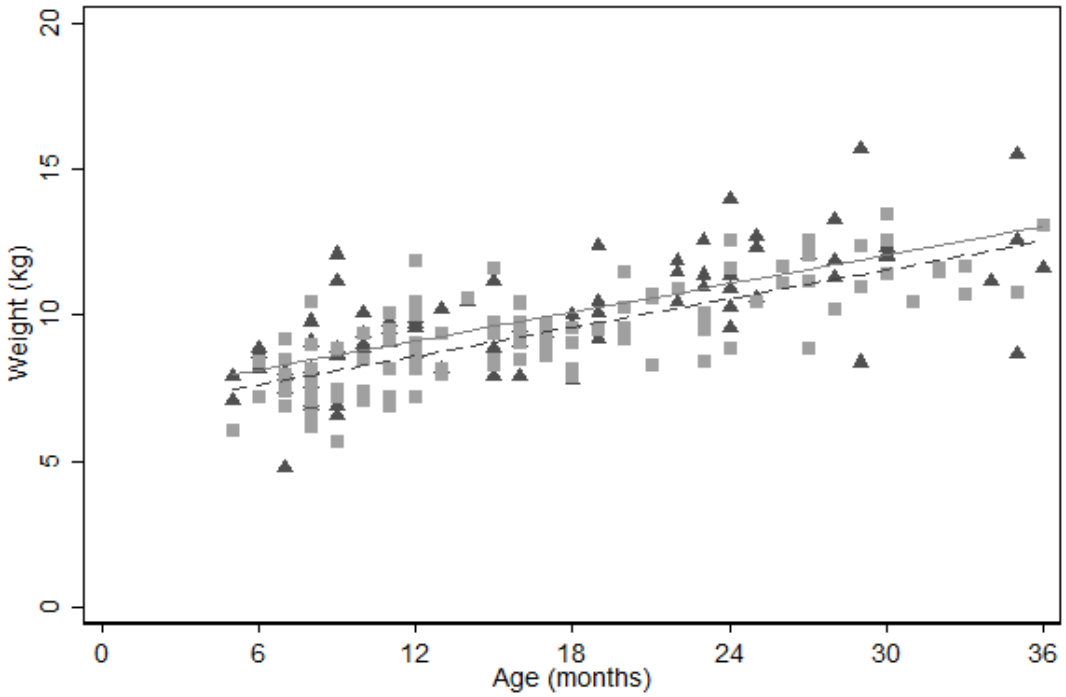


Figure 1: Data and fitted values from a regression model relating age and gender to data from the Gambian cross-sectional survey. For male children data points shown as triangles and fitted values linked by a solid line. For female children data points shown as squares and fitted values linked by a dashed line.

Exercise: Provide a full interpretation of the results of the above output of the ANCOVA model relating age and sex of infants to their weight in the Gambian cross-sectional survey.

4.6 Changes in partial regression coefficients

One of the most important challenges facing a data analyst using linear regression models is to understand the reasons why estimates and the inferences made from them change when one or more additional predictor variables are added to a regression model. Here we focus on changes in partial regression coefficients, returning to other changes (such as in the standard errors) in regression 5.

As an introduction we consider the impact of adding a second predictor variable to a simple regression model with a single predictor variable. We will use the following notation for the models of interest.

Model 1: $y_i = \alpha^* + \beta_1^* x_{1i} + \varepsilon_i^*$
Model 2: $y_i = \alpha\ + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

For example we might consider *Y* to be a score on an anxiety questionnaire in a cross-sectional survey of young adults and X_1 being hours of social media use in the last week and X_2 being units of alcohol drunk in the last week.

In general we anticipate that $\beta_1^* \neq \beta_1$ since the interpretation of β_1^* and β_1 is fundamentally different. β_1^* quantifies the effect of a unit change in X_1 on *Y* without reference to X_2 whilst β_1 quantifies the effect of a unit change in X_1 on *Y* holding X_2 constant. Specifically if a unit change in X_1 is associated with some change in X_2 and this change in X_2 is itself associated with a change in *Y* then β_1^* and β_1 will be different.

This difference between the **crude** (β_1^*) and **adjusted** (β_1) effects can be quantified for a <u>linear regression model</u>. To do this we specify a third model relating X_2 to X_1:

Model 3: $x_{2i} = \gamma + \delta_1 x_{1i} + \omega_i$

Substituting this representation of x_{2i} into Model 2:

$$\begin{aligned} y_i &= \alpha\ + \beta_1 x_{1i} + \beta_2(\gamma + \delta_1 x_{1i} + \omega_i\ ) + \varepsilon_i \\ &= \alpha\ + \beta_2\gamma + (\beta_1 + \beta_2\delta_1)x_{1i} + \beta_2\omega_i\ \ + \varepsilon_i \end{aligned}$$

Comparing the coefficients for x_{1i} in Model 2 and the above, we find:

$$\beta_1^* = \beta_1 + \beta_2\delta_1$$

This means that that difference between the **crude** (β_1^*) and **adjusted** (β_1) effects depends upon two factors.

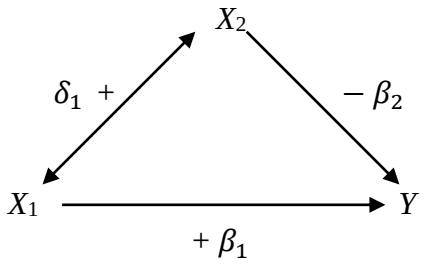
i) The sign and magnitude of β_2.
ii) The extent to which X_1 and X_2 are associated with each other (which we have quantified by δ_1 in Model 3).

One very important point to appreciate is that the change in the partial regression coefficient (β_1) is predicted by the sign and magnitude of the magnitude of the regression coefficients in the <u>adjusted</u> model that included both X_1 and X_2.

There are three potential scenarios that may result from these associations and we will discuss each of them in turn under the situation where the adjusted effect of X_1 on *Y* is positive ($\beta_1 > 0$).

4.6.1 Scenario 1: $\boldsymbol{\beta_1 > \beta_1^*}$

This scenario occurs if $\beta_2\delta_1$ is negative, which will be when one of β_2 and δ_1 is negative and the other is positive. One example would be as follows.



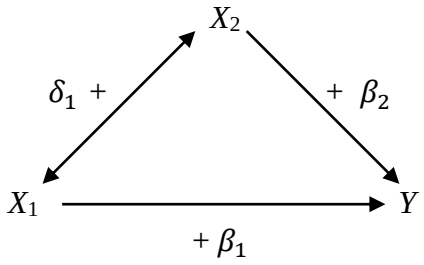
Increasing X_1 (holding X_2 constant) is associated with an increase in Y. However, an increase X_1 is also associated with an increase in X_2, which is itself associated with a concomitant decrease in Y. This means if we do not hold X_2 constant then the total increase in Y per unit increase in X_1 is less than it would be were X_2 held constant.

The extent to which the crude effect is less than the adjusted effect is determined by the magnitude of the various associations. The larger the magnitude of β_2 and δ_1, the greater the difference between β_1^* and β_1.

When the adjusted effect of X_1 on Y is positive (i.e. $\beta_1 > 0$) then the unadjusted coefficient (β_1^*) may be closer to zero but in the same direction. However, if $\beta_2\delta_1$ is large enough, the direction of the adjusted coefficient may be different to that of the crude regression coefficient. For the above example it might be found that β_1^* was negative even though β_1 was positive.

4.6.2 Scenario 2: $\boldsymbol{\beta_1 < \beta_1^*}$

This scenario occurs if $\beta_2\delta_1$ is positive, which will be when both β_1 and δ_1 are positive OR both β_1 and δ_1 are negative. One example would be as follows.

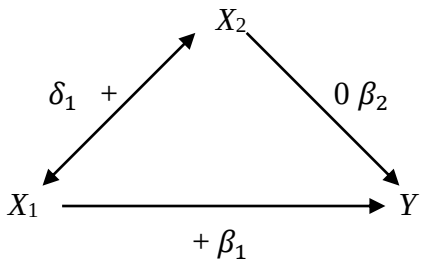


Increasing X_1 (holding X_2 constant) is associated with an increase in Y. However, increasing X_1 is associated with an increase in X_2 which is itself associated with an increase in Y. So, when we do not hold X_2 constant the net increase in Y per unit increase in X_1 is greater than it would be were X_2 held constant.

As before, when β_2 and δ_1 are larger in magnitude, there is a greater difference between β_1^* and β_1. If the adjusted effect of X_1 on Y is positive (i.e. $\beta_1 > 0$), as in this example, then the adjusted coefficient (β_1) will be closer to zero but in the same direction as the unadjusted coefficient (β_1^*).

4.6.3 Scenario 3 $\boldsymbol{\beta_1 = \beta_1^*}$

This scenario occurs if $\beta_2\delta_1$ is zero, which will be when either β_1 or δ_1 are zero. For example:



Increasing X_1 (holding X_2 constant) is associated with an increase in Y. Increasing X_1 is associated with an increase in X_2, but this is not associated with any change in Y. So when we do not hold X_2 constant, the net increase in Y per unit increase in X_1 is identical to that if X_2 were held constant.

Exercise: Repeat the scenarios 1-3 under the situation where the adjusted effect of X_1 on Y is negative (i.e. $\beta_1 < 0$). When might the unadjusted effect be positive even though the adjusted effect is negative? Under which scenarios would the adjusted coefficient be closer to zero than the unadjusted coefficient?

4.6.4 Example

The following results are from the REPAIR study of live donor kidney transplant, introduced in section 4.3. Stata has been used to relate recipient's glomerular filtration rate (GFR) at 12 months after transplant (`gfr_iohexol`) to donor's age in years (`d_age_b`) and recipients age in years (`r_age_b`) at the start of the study.

```
. regress gfr_iohexol d_age r_age_b

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     d_age_b |  -.5118793   .0690202    -7.42   0.000    -.6476733   -.3760853
     r_age_b |  -.0130273    .057206    -0.23   0.820    -.1255774    .0995227
       _cons |   84.93954   3.872341    21.93   0.000      77.3209    92.55819
------------------------------------------------------------------------------

. corr  d_age r_age_b if gfr_iohexol!=.
(obs=321)
             |  d_age_b  r_age_b
-------------+------------------
     d_age_b |   1.0000
     r_age_b |   0.2482   1.0000


. regress gfr_iohexol d_age_b

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     d_age_b |  -.5157799   .0667616    -7.73   0.000    -.6471287   -.3844312
       _cons |   84.52499   3.412683    24.77   0.000     77.81078     91.2392
------------------------------------------------------------------------------
```

```
. regress gfr_iohexol r_age_b

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     r_age_b |  -.1183151   .0599238    -1.97   0.049    -.2362108   -.0004194
       _cons |   64.48641   2.939435    21.94   0.000     58.70328    70.26954
------------------------------------------------------------------------------
```

We can see that the estimated effect of recipient age on recipient GFR is much larger in the simple regression model than in model that include both recipient age and donor age. This is due to the association between recipient age and donor age. Patients with older kidney donors tend, on average, to be older themselves. Those with older donors also tend to have lower kidney function (GFR). Therefore, in the unadjusted analysis the negative slope of the relationship between recipient age and kidney function is exaggerated because those patients who are older tend to have older donors, and having an older donor is associated with lower kidney function.

After adjusting for donor age there is little association between recipient age and GFR. This means that the association between donor age and recipient GFR is reasonably similar whether the model includes recipient age or not.

4.7 Confounding

The epidemiological term **confounding** is often used to explain why crude and adjusted regression coefficients differ from one another. Confounding occurs when there are predictors, which might be observed or unobserved, that obscure the true causal association between a predictor variable of interest and the outcome. This will happen when the exposure and outcome of interest share common causes. For example the effect of exposure to air pollution on lung function could be confounded if higher levels of exercise were a common cause of both better lung function and higher exposure to pollution. In this case we can call exercise a **confounder** of the association.

In order to make **causal statements** from the results of a linear regression model, such as "higher exposure to air pollution causes reductions in lung function", it is necessary to identify a set of variables that when adjusted for remove any influence of confounding. However, judging which variables we need to adjust for to control for confounding is not straightforward. We discuss some of the challenges below.

4.7.1 Assessing confounding in observational studies

In order for a variable to confound an association of interest, it must be associated with both the predictor and outcome in the population. Therefore, it may seem intuitive that we could examine these associations in our data to determine if a variable is a confounder or not. While this can be useful, this approach has some major limitations.

One issue is that a variable may be associated with both predictor and outcome, but not be a confounder. A common situation where is occurs is if a variable is on the **causal pathway** between the predictor of interest and the dependent variable. For example, if we were examining the effects of a blood pressure lowering drug on aortic aneurysm size then we would

find that blood pressure was associated with both the exposure (taking the drug) and the outcome (aortic aneurysm size). Nevertheless, blood pressure would not be confounding the association. We might say that some of the effect of the treatment on aortic aneurysm size was **mediated** through the effects of treatment on blood pressure.

We should also be aware that confounding may occur due to unobserved variables, and it is impossible to test for this using the data. For these reasons, we cannot use analysis of the data to rule out confounding.

To help in deciding how to model an association, subject matter knowledge can be used to construct a **causal diagram**, which sets out the assumed causal relationships between the variables. We will not discuss use of causal diagrams in detail here but the papers by Greenland et al (1999) and Hernán et al (2002) are suggested as an introduction to this important topic. Causal diagrams can be helpful in identifying which variables we would need to control for in order to obtain an unbiased estimate of a causal effect and whether this is possible given the data we have collected.

- Greenland et al. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10:37-48.
- Hernán et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 2002; 155: 176-184.

Although we cannot use the data to determine if confounding is present or not, we can examine the associations between variables to inform our proposed model. In the REPAIR example, we found no association between recipient age and GFR in the adjusted model, which suggests that it was not necessary to control for recipient age to remove confounding of the association between donor age and GFR. One important point to note here is that we should not rely results of statistical tests to detect confounding. For example, in a small dataset we may find that a variable does not have a "statistically significant" association with the outcome of interest, but it could still be an important confounder.

The need to base our modelling decisions on assumptions that cannot be determined from the data alone means that in the vast majority of situations we cannot be completely certain whether an observed association could be confounded (or not). One exception to this is where we are analysing data from randomised controlled trials.

4.7.2 An exception: randomised controlled trials

Randomised controlled trials are an exception where (if the trial was well conducted) we do know there is no confounding of the treatment outcome association. Due to randomisation, all baseline variables (either observed or unobserved) are unrelated to treatment allocation. We therefore know the value of the population parameters relating each baseline variable to treatment (i.e. $\delta = 0$ for any model relating a baseline variable to the treatment). Therefore, we can be certain there is no confounding of the association between treatment allocation and the outcome. Of course we might see chance imbalances in observed baseline characteristics in a randomised trial, and so the parameter estimates would suggest an association between this baseline characteristic and the treatment allocation. However, this should not really be called confounding. This illustrates the confusion that can arise through defining confounding in terms of the behaviour of parameter estimates rather than that of parameters.

Regression 5: Multivariable Models Continued

5.1 Objectives

By the end of this session students will be able to:

- Formulate statistical models using matrix and 'non-matrix' notation.
- Fit and interpret multiple linear regression models using Stata.
- Perform and interpret global and partial F-tests both by hand calculation and using Stata.
- Explain the reasons for changes in standard errors when new predictor variables are added to a regression model.

5.2 Introduction

In Regression 4 linear regression models with two predictor variables were introduced. In this session we extend this framework to models with three or more predictor variables. We also introduce the hypothesis tests used in linear regression. The most general tests are F-tests, although, as with simple linear regression, these are equivalent to t-tests when only a single parameter is to be tested.

Conceptually there is little that is new here; as one might expect from the models introduced in Regression 4 the aim of a model with more than two predictor variables is to understand how each of them relates to the dependent variable whilst holding all of the others constant. These types of relationship can be described as **conditional** relationships, although when using this term it is always important to specify what is being conditioned on. For example in the Gambian cross-sectional data we might wish to investigate the relationship between weight and length, conditional on age and gender.

Visualising the relationships becomes progressively more complex as the number of predictor variables increases. In simple linear regression we can visualise the model as a line of best fit in two dimensional space. With two predictors the fitted values are a plane in three dimensional space. With three predictors we can introduce 'time' as a third predictor variable and the fitted values are then described by a plane moving with constant speed through three dimensional space. Visualising models with more predictors than this is difficult! A consequence of this is that the methods introduced in Regression 6 for checking assumptions are more important for models with a large number of predictor variables.

In the final part of the session we will return to the changes that take place when a new predictor variable is added to a regression model. In particular we focus on changes in standard errors and the important topic of **collinearity**.

5.3 Linear regression models in matrix and 'non-matrix' notation.

5.3.1 Formulation of the model

Suppose we wish to relate a dependent variable (Y) to p predictor variables ($X_1,\ldots, X_p$). The linear regression model here is as follows.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \text{ with } \varepsilon_i \sim NID(0, \sigma^2) \quad (1)$$

Where,

y_i = value of the dependent variable for the ith participant
x_{pi} = value of the pth predictor variable for the ith participant

This relationship can also be expressed using matrix algebra.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad where\ \varepsilon \sim N(0, \boldsymbol{I}\sigma^2) \quad (2)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & . & x_{p1} \\ 1 & x_{12} & x_{22} & . & x_{p2} \\ 1 & . & . & . & . \\ 1 & . & . & . & . \\ 1 & x_{1n} & x_{2n} & . & x_{pn} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

In this formulation $\boldsymbol{X}$ is an $n \times (p + 1)$ matrix, $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are vectors of length n whilst $\boldsymbol{\beta}$ is a vector of length $(p + 1)$ whose first element is α (sometimes written as β_0). These are the most commonly adopted ways of writing these equations. One potentially confusing aspect of this is that x_{ij} in equation (1) is the (j+1, i)th element of $\boldsymbol{X}$ in equation (2). This reversal of i and j means that some authors prefer to define the matrix termed $\boldsymbol{X}$ in equation (2) as $\boldsymbol{X}'$ (the transpose of $\boldsymbol{X}$).

The residuals are assumed to follow a **multivariate normal distribution** with variance-covariance matrix equal to σ^2 multiplied by the identity matrix. This is equivalent to assuming that residuals are normally and independently (implying that the covariances in the multivariate formulation are all zero) distributed with constant variance σ^2.

5.3.2 Interpretation of the parameters

The parameters in the model are as follows:

α is the intercept. It is the expectation of Y when **all the X_j's are zero**.

β_j is the increase in the expectation of Y for a **1 unit increase in X_j with all the other predictor variables held constant**.

As before the β_j's are often termed **partial** regression coefficients. Each one measures the effect of one predictor variable **controlled** (or **adjusted**) for all of the others.

5.3.3 Least Squares Estimation

As with models presented earlier estimates of regression parameters are obtained by minimising the residual sum of squares.

$$SS_{RES} = \sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$
$$= \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_p x_{pi}\right)^2 \quad (3)$$

The closed form solution, obtained by solving the (p + 1) simultaneous equations that result from setting the partial derivatives of (3) with respect to each parameter estimate to zero, can be written succinctly using matrix notation.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (4)$$

$\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Its distribution is as follows:

$$\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\sigma}^2) \quad (5)$$

Equation (5) expresses the fact that the elements of $\widehat{\boldsymbol{\beta}}$ follow a multivariate normal distribution whose variances and covariances are given by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\sigma}^2$.

It can also be shown that the following is an unbiased estimator for σ^2.

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon_i}^2}{(n-(p+1))} = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{pi}\right)^2 / (n-(p+1)) \quad (6)$$

5.3.4 Expected values of the dependent variable

Formulae for the vector of expected values ($\hat{\boldsymbol{Y}}$) follow from equation (4).

$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}$ where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (7)

The ($n \times n$) matrix $\boldsymbol{P}$ is a key matrix in multiple regression. It is often referred to as the 'hat' matrix as it 'projects' the observed $\boldsymbol{Y}$ onto their fitted values (Y 'hats'). Note also that the ith diagonal element of $\boldsymbol{P}$ is the **leverage** for that observation (Regression 6). One important property of $\boldsymbol{P}$ is that $\boldsymbol{P}^2 = (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{P}$. This implies that projecting a set of points onto fitted values and then projecting the fitted values onto new fitted values does not result in any further changes). It follows from this result and from (7) that the variance-covariance matrix of the fitted values is given by:

$$Var(\widehat{\boldsymbol{Y}}) = \boldsymbol{P}\sigma^2 \quad (8)$$

5.3.5 Residuals

(Observed) residuals are defined to be the distance between the observed and fitted values. It follows from the results given above that the observed residuals can be written as:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{PY} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \qquad (9)$$

Using simple matrix multiplication it can be shown that the variance-covariance matrix of the residuals is given by:

$$Var(\hat{\boldsymbol{\varepsilon}}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2 \qquad (10)$$

Since $\boldsymbol{P}$ is not, in general, a diagonal matrix this demonstrates that observed residuals are not independent. Further, since the diagonal elements of $\boldsymbol{P}$ are not (in general) all equal the variance of the residuals are not all the same, motivating consideration of **standardised** residuals as discussed in Regression 6.

5.4 Analysis of Variance and *F*-tests: General Formulation

5.4.1 Coefficient of Determination and residual variance in multiple regression

As with a simple regression model it can be shown that the corrected sums of squares can be partitioned into two components: the sum of squares explained by the regression model and the unexplained residual sum of squares. As in Regression 3 (equation (1)) the key equation is:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$SS_{yy} \quad = \quad SS_{REG} \quad + \quad SS_{RES} \qquad (11)$$

As with simple linear regression the coefficient of determination is defined as follows.

$$R^2 = \frac{SS_{yy} - SS_{RES}}{SS_{yy}} = 1 - \frac{SS_{RES}}{SS_{yy}} = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (12)$$

R^2 is interpretable as the proportion of the variability in the dependent variable explained by the model.

When a new predictor variable is added to a regression model the residual sum of squares will never increase. This is because the estimates of the parameters are those that minimise the residual sum of squares and one 'candidate' set of parameters is that in which the partial regression coefficient for the 'new' predictor variable is zero (*i.e.* the original simpler model).

Hence R^2 will never decrease when a new predictor variable is added to the model (provided, of course, that the number and identity of the observations stays the same). This means that R^2 cannot be directly used to compare the fits of models with different numbers of predictors. So-called **adjusted $\boldsymbol{R^2}$s** can be calculated to address this issue (and are reported by many software packages, including Stata), but they are not widely used and will not be considered in this module.

Although the residual sum of squares can never increase when a new variable is added to a regression model the estimated residual variance can increase (because this is estimated by dividing the residual sum of squares by the number of observations minus the number of estimated parameters).

5.4.2 The analysis of variance table

As with the simple linear regression models, the i) total (corrected), ii) residual and iii) model sums of squares (and the corresponding mean squares) from a model with p predictor variables can be conveniently displayed in an Analysis of Variance table as shown in table 1. The elements of the ANOVA table can be used to perform tests of null hypotheses relating to the slope parameters in the model. These tests are the global and partial F-tests.

Source of variation due to	Sum of Squares	Degrees of Freedom	Mean Sum of Squares
Regression (model)	SS_{REG}	p	$MS_{REG} = SS_{REG} / p$
Residual	SS_{RES}	$n - (p + 1)$	$MS_{RES} = SS_{RES}/(n - (p + 1))$
Total	SS_{yy}	$n - 1$	$SS_{yy}/(n - 1)$

Table 1: Analysis of Variance table for a linear regression model with p predictor variables

5.4.3 The Global *F*-test

The global F-test tests the null hypothesis that all the slope parameters (β's) in equation (1) are equal to zero. It is important to understand that the alternative hypothesis is not that all of the β's are non-zero, only that at least one of them is non-zero. There are strong parallels here with the global F-test for one-way ANOVA models introduced in Regression 3 (section 3.4.2). The appropriate test statistic here is the ratio of the mean sum of squares explained by the model to the mean residual sum of squares. Under the null hypothesis this test statistic follows an F-distribution. Specifically

$$\text{Under } H_0 \; F = \frac{MS_{REG}}{MS_{RES}} \sim F_{p,(n-(p+1))} \qquad (13)$$

Hence under H_0 the expectation of F is close to 1 (actually $1 + (2/(n - p - 3))$). Further, under the alternative hypothesis the expectation of F will be greater than this. Hence a comparison of F with the upper tail of the $F_{p,(n-(p+1))}$ distribution yields a p-value for testing the null hypothesis. This F statistic and its associated p-value are reported by Stata as shown overleaf for the data from the Gambian cross-sectional study considered in Regression 4.

```
. regress wt age len                                    Model 3:

      Source |       SS       df       MS              Number of obs =     190
-------------+------------------------------           F(  2,   187) =  270.84
       Model |   493.57848     2   246.78924            Prob > F      =  0.0000
    Residual |  170.391278   187  .911183307            R-squared     =  0.7434
-------------+------------------------------           Adj R-squared =  0.7406
       Total |  663.969759   189  3.51306751            Root MSE      =  .95456

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0112598   .0167512    -0.67   0.502    -.0443055    .0217859
         len |   .2371289   .0195165    12.15   0.000     .1986281    .2756296
       _cons |  -8.351244   1.259968    -6.63   0.000    -10.83682   -5.865667
------------------------------------------------------------------------------
```

Exercise: i) Show how the *F*-statistic is calculated from the elements of the ANOVA table.
ii) What conclusions can be drawn from the result of the *F*-test?

5.4.4 The Partial *F*-test

The global *F*-test is a **joint** test of the statistical significance of **all** the slope parameters in a linear regression model. However often we wish to compare the fit of a complex model (model B with p predictor variables) with a simpler model (model A) in which **some** ($k < p$) of these slope parameters are zero. The appropriate test here is the partial *F*-test.

The key to the partial *F*-test is the construction of an Analysis of Variance table that partitions the sum of squares explained by the complex model into that explained by the simple model and the **extra sum of squares** only explained by the complex model. Using the notation that SS_{REG_B} denotes the sum of squares explained by the complex model, whilst SS_{REG_A} denotes the sum of squares explained by the simpler model, the ANOVA table is as shown in Table 2.

As with earlier models since $SS_{REG\_B} + SS_{RES\_B} = S_{yy} = SS_{REG\_A} + SS_{RES\_A}$ it follows that the extra sum of squares can be expressed as either the difference in the explained sum of squares ($SS_{REG\_B} - SS_{REG\_A}$) or as the difference in the residual sum of squares ($SS_{RES\_A} - SS_{RES\_B}$) from the two models.

Variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares
Explained by model A	SS_{REG_A}	$p - k$	$MS_{REG\_A} = SS_{REG\_A} / (p - k)$
Extra explained by model B	$SS_{REG\_B} - SS_{REG\_A}$	k	$(SS_{REG\_B} - SS_{REG\_A}) / k$
Residual from model B	SS_{RES_B}	$n - (p + 1)$	$MS_{RES} = SS_{RES} / (n - (p + 1))$
Total	SS_{yy}	$n - 1$	$SS_{yy} / (n - 1)$

Table 2: Analysis of Variance table comparing the fit of a model (B) with p predictor variables with that of one (model A) with $(p - k)$ predictor variables.

The partial *F*-test tests the null hypothesis that all of the slope parameters included in model B but omitted from model A are equal to zero. If (without loss of generality) the last *k* predictor variables are omitted from model A then the null hypothesis is H_0: $\beta_{p-k+1} = \beta_{p-k+2} = \ldots \beta_p = 0$ and the alternative hypothesis is that at least one of these *k* parameters is non-zero. The appropriate test statistic here is the ratio of the extra sums of squares explained by the complex model divided by the number of additional parameters in the more complex model to the mean residual sum of squares from the complex model. Under the null hypothesis this test statistic follows an *F*-distribution. Specifically

$$\text{Under } H_0 \; F = \frac{(SS_{REG_B} - SS_{REG_A})/k}{MS_{RES_B}} \sim F_{k,(n-(p+1))} \qquad (14)$$

Hence under H_0 the expectation of *F* is close to 1 (actually $1 + (2/(n - p - 3))$). Further, under the alternative hypothesis the expectation of *F* will be greater than this. Hence a comparison of *F* with the upper tail of the $F_{k,(n-(p+1))}$ distribution yields a *p*-value for testing the null hypothesis. This calculation is illustrated below with reference to testing the null hypothesis that both the partial regression coefficients relating age and gender to weight are zero in a model that additionally relates weight to length in the Gambian cross-sectional data.

Model A:

```
. regress wt len

      Source |       SS       df       MS          Number of obs =     190
-------------+------------------------------       F(  1,   188) =  542.82
       Model |  493.166786     1  493.166786       Prob > F      =  0.0000
    Residual |  170.802973   188  .908526451       R-squared     =  0.7428
-------------+------------------------------       Adj R-squared =  0.7414
       Total |  663.969759   189  3.51306751       Root MSE      =  .95317

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         len |   .2257467   .0096893    23.30   0.000     .2066329    .2448605
       _cons |  -7.669441   .7463556   -10.28   0.000    -9.141748   -6.197133
------------------------------------------------------------------------------
```

Model B:

```
. regress wt len age i.sex

      Source |       SS       df       MS          Number of obs =     190
-------------+------------------------------       F(  3,   186) =  182.08
       Model |  495.311567     3  165.103856       Prob > F      =  0.0000
    Residual |  168.658192   186  .906764474       R-squared     =  0.7460
-------------+------------------------------       Adj R-squared =  0.7419
       Total |  663.969759   189  3.51306751       Root MSE      =  .95224

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         len |   .2317997    .019847    11.68   0.000     .1926454    .2709539
         age |  -.0077959   .0168974    -0.46   0.645     -.041131    .0255392
       2.sex |  -.1964758   .1421171    -1.38   0.168    -.4768443    .0838928
       _cons |  -7.890676   1.300309    -6.07   0.000    -10.45593   -5.325426
------------------------------------------------------------------------------
```

$$F = \frac{(SS_{REG_B} - SS_{REG_A})/k}{MS_{RES_B}} = \frac{(495.312 - 493.167)/2}{0.907} = 1.18$$

Comparison of F with $F_{2,186}$ gives a p-value of 0.31 and hence we have no evidence against the null hypothesis of no association between weight and either age or gender after adjusting for length.

The partial F-test can be carried out using the **`test`** command in Stata. After fitting model B the following command could be used.

```
. test age 2.sex

 ( 1)  age = 0
 ( 2)  2.sex = 0

       F(  2,   186) =    1.18
            Prob > F =    0.3088
```

Exercise: Interpret each of the estimated partial regression coefficients in model B.

The example above illustrates how the partial F-test can be used to compare two nested models where the respective number of predictors differ by two. The **`test`** command in Stata can also be used to carry out global F-tests. For example in model B above the global F-test can be carried out as follows. However this is of little value since it merely duplicates the result reported routinely in the Stata output.

```
. test len age 2.sex

 ( 1)  len = 0
 ( 2)  age = 0
 ( 3)  2.sex = 0

       F(  3,   186) =  182.08
            Prob > F =    0.0000
```

5.4.5 Hypothesis tests for a single parameter

The partial F-test can also be used to compare the fits of two models which differ only in the inclusion or exclusion of a single parameter. However in this situation, the F and t tests of the null hypothesis that the parameter in question is zero are equivalent (with $F = t^2$) giving identical p-values and hence can be considered to be the same statistical test.

As a reminder the Wald statistic to test the null hypothesis H_0:β=0 against the alternative H_1:$\beta \neq 0$ is:

$$t = (\hat{\beta} - 0)/SE(\hat{\beta})$$

This test statistic follows a t-distribution with (n-(p + 1)) degrees of freedom (rather than a z-distribution) if H_0 is true. For example, for model B the degrees of freedom for the t test would be 186 (190-(3+1)).

A t distribution with (n-(p + 1)) degrees of freedom is also used when constructing confidence intervals for each coefficient. For example, the 95% confidence interval for the parameter β is given by:

$$\hat{\beta} \pm t_{n-(p+1),\, 0.975} SE(\hat{\beta})$$

The equivalence of *F* and *t* tests for a single parameter is illustrated below.

```
. regress wt len age i.sex                                    Model B:

      Source |       SS       df       MS       Number of obs =     190
-------------+------------------------------     F(  3,   186) =  182.08
       Model |  495.311567     3  165.103856     Prob > F      =  0.0000
    Residual |  168.658192   186  .906764474     R-squared     =  0.7460
-------------+------------------------------     Adj R-squared =  0.7419
       Total |  663.969759   189  3.51306751     Root MSE      =  .95224

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         len |   .2317997    .019847    11.68   0.000     .1926454    .2709539
         age |  -.0077959   .0168974    -0.46   0.645     -.041131    .0255392
       2.sex |  -.1964758   .1421171    -1.38   0.168    -.4768443    .0838928
       _cons |  -7.890676   1.300309    -6.07   0.000    -10.45593   -5.325426
------------------------------------------------------------------------------

. test 2.sex

 ( 1)  2.sex = 0

       F(  1,   186) =    1.91
            Prob > F =    0.1685
```

The *F*-statistic testing the null hypothesis of no difference in weight between males and females after adjusting for age and length (1.91) is the square of the t-statistic (1.38) and the p-values are the same (although Stata reports one to 3 decimal places and the other to four).

In this session we have focussed on the mechanics of hypothesis testing. When building a multiple regression model to describe a potentially complex relationship between a dependent variable and many potential predictor variables the strategy of analysis is very important. As an illustration, if we were analysing the cross-sectional data from the Gambia, we would need to consider whether is it best to relate weight to age, gender and length simultaneously, and then to jointly test the statistical significance of age and gender, or whether we should test individual predictor variables and potentially remove them one at a time. The subject of model building strategies is returned to in term 2.

5.5 Impact of adding a new predictor variable to a regression model

As discussed in Regression 4 one of the most important challenges facing a data analyst using linear regression models is to understand the reasons why estimates and the inferences made from them change when one or more additional predictor variables are added to a regression model. In general when a new predictor variable is added to a regression model all of the following change:

(a) Estimated regression coefficients for variables in the simpler model change.
(b) The variances of these partial regression coefficients change.
(c) The statistical significance of these partial regression coefficients change.
(d) Predicted values change.
(e) R^2 increases.

Reasons for the changes in estimated partial regression coefficients were discussed in detail in Regression 4. Here we focus on the other changes, starting with the changes in the variances of the partial regression coefficients.

5.5.1 Changes in variance of estimators of partial regression coefficients

In 5.3.3 (equation (5)) it was stated that the elements of the vector of partial regression coefficients $\widehat{\boldsymbol{\beta}}$ follow a multivariate normal distribution whose variances and covariances are given by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\sigma}^2$. This illustrates that the following are the determinants of the variance of the estimator of a particular partial regression coefficient.

(a) The residual variance.
(b) The sample size.
(c) The covariances between the predictor variable in question and the others.

The fact that the first two of these are determinants of precision is intuitive. The dependency on the covariances follows from the fact that the precision of partial regression coefficients depends on the spread of values taken by the predictor variables. In simple linear regression the larger the spread of values (as measured by the standard deviation), the more precise the estimate of the slope. In multiple regression it is the variability of the predictor variable whilst the other predictor variables are held constant that is relevant and this will be smaller if the predictor variables are highly correlated one with another than if they are not.

The fact that the precision of parameter estimates depends both on the residual variance and on the covariance structure of the predictor variables means that when a new predictor variable is added to a regression model the variances of regression coefficients for the predictor variables already in the model can either increase or decrease. If the predictor variable that is added to the model explains a large part of the residual variance then this will tend to increase precision. However if the new predictor is highly correlated with a particular existing predictor variable then the variance of that partial regression coefficient may well increase. When highly correlated predictor variables are included in regression models the effects are often described as arising through collinearity, a subject we return to in section 5.5.5.

5.5.2 Changes in p-values of partial regression coefficients

When a new variable is added to a regression model the p-values for the partial regression coefficients relating to the predictor variables already in the model change. This can occur either because of changes in the estimated partial regression coefficient and/or because of changes in the variance of the estimate. In seeking to interpret changes in the p-value it is best to first seek to understand these component changes, rather than to try and directly understand the change in the p-value.

As an illustration suppose that the coefficient for a particular predictor variable (X_1) has a p-value of 0.01 in simple linear regression for a dependent variable (Y), but in a multiple regression model when a second predictor variable (X_2) is added the p-value changes to p=0.24. Interpretation of this result is very different if i) the partial regression coefficient representing the effect of X_1 is much reduced when X_2 is added to the model or ii) the partial regression coefficient remains materially unaltered but its variance is increased.

5.5.3 Changes in predicted values

Fitted values change when a new predictor variable is added to a model. Typically the introduction of the new variable has less impact on the fitted value for a particular observation than it does on estimates of partial regression coefficients, unless the observation in question has a particularly atypical value for the new predictor conditional on the other predictor variables in the model.

5.5.4 Changes in R^2

As explained in section 5.4.1 the coefficient of determination always increases when a new variable is added to a regression model.

5.5.5 Collinearity

Strictly collinearity occurs when one explanatory variable (say X_1) is an exact linear combination of the others. In this situation it is not possible to estimate the partial regression coefficients, since it is impossible to change X_1 whilst holding all of the other predictor variables constant. In matrix notation, $\boldsymbol{X}$ is said to be **singular** and $(\boldsymbol{X'X})^{-1}$ does not exist.

An example of exact collinearity occurs if we convert a K-level categorical variable into K dummy variables and try to include all of these (and the intercept) as predictor variables in a multiple regression model. Since we can consider the intercept term in a linear regression model to be the partial regression coefficient associated with a predictor variable whose value is 1 for all observations it follows that the sum of the dummy variables minus this constant predictor variable will be zero for all observations. This indicates that each of the dummy variables is an exact linear combination of the other predictor variables and it will not be possible to fit the model without dropping one of them (as will happen automatically if a statistical package such as Stata is used to fit the model).

The term collinearity is also used when two or more predictor variables are highly correlated (and so $\boldsymbol{X}$ in matrix notation is nearly singular). Collinearity introduces instability into a regression model. The following are some manifestations of this.

(a) Estimated variances of partial regression coefficients are large.
(b) There are surprisingly large (in magnitude) estimated partial regression coefficients.
(c) Known important variables have surprisingly small and non-statistically significant estimated partial regression coefficients.

Adding a new predictor variable that is highly correlated with existing ones may cause any or all of the above effects. However it rarely has a major impact on predicted values as mentioned in 5.5.3 above.

To avoid collinearity it is necessary either to remove one of the highly correlated variables or to transform one of them. Examples of such transformations include

(a) Instead of using systolic and diastolic blood pressure as collinear predictor variables, use diastolic blood pressure and (systolic - diastolic blood pressure).

(b) Instead of using height and weight as predictor variables, use height and body mass index (weight/height^2), which are less highly correlated.

(c) When fitting a quadratic regression model use X and $(X - \bar{X})^2$, rather than X and X^2 as predictor variables.

Exercise: Which of the above changes ((a) – (c)) will alter the fitted values from the regression model and which will not? Give reasons for your answer.

Regression 6: Checking the Assumptions of the Linear Model

6.1 Objectives

By the end of this session students will be able to:

- Apply a range of graphical techniques to investigate the assumptions of the normal regression model.
- Explain how to modify the linear regression model to allow for a non-linear association between a predictor and the dependent variable.
- Calculate and interpret the statistics used to describe the impact that individual points have on the results of a regression analysis.

6.2 Introduction

Like most other statistical procedures the linear regression model makes certain assumptions and strictly all inferences made from a model are contingent on these assumptions being correct. It is therefore important that we have statistical techniques to investigate these assumptions and several such techniques are introduced in this session.

The role of assumptions in statistical procedures has been considered in some detail in Analytical Techniques 6. As explained there, in practice it is rare for all the assumptions of a procedure to hold exactly. Often we may have evidence in the data, or suspicions raised by knowledge of the type of data under consideration, that the assumptions made by the model do not hold. This does not always mean that results from using the linear model in question should necessarily be disregarded, since we know that in many settings many statistical procedures are robust to departures from assumptions. Nonetheless it is always a good idea to first try and establish to what extent assumptions hold and then to consider whether the procedures used can be adapted to improve the extent to which assumptions hold. If adaptations cannot be made then it is necessary to consider to what extent the results of an analysis where assumptions are violated can be 'broadly trusted' through consideration of the robustness of the technique.

In this session we shall largely focus on techniques for the **identification** of problems. Some pointers will be given to adaptations and alternative techniques that can be used when assumptions are violated. Issues of robustness will not be considered in much detail. However at the outset it is worth recalling (from Analytical Techniques 6) that, broadly speaking, the central limit theorem implies that departures from assumptions are less important for large datasets than for small ones.

6.3 Assumptions of the linear regression model

The assumptions made by the linear regression model are as follows.

1) There is a **linear** relationship between the dependent variable (Y) and each of the predictor variables. Here we are contrasting a linear relationship with a non-linear relationship, not with no relationship. A model in which one of the regression coefficients is zero can satisfy the assumptions of linear regression.

2) The observations y_i are **independent**.

3) The **true** residual variance is **constant**. *i.e.* the scatter of points around the true regression line has the same variance, irrespective of the value of x_i. This feature is termed **homoscedasticity** (homogeneity of the residual variance) with its converse being **heteroscedasticity** (heterogeneity of the residual variance).

4) The **true** residuals follow a normal distribution. In fact linear regression models can be fitted without this assumption. However, when we use the parametric framework introduced in this module we do require this assumption to provide valid inference (p-values, confidence intervals, etc.)

In this session we will focus on the first, third and fourth of these assumptions. Violations of the second assumption (independence) are often more apparent from the context of a study than from the data itself. For example, if we carry out a study in which the blood pressures of 100 people are each measured twice, and then treat the 200 measurements as independent in the statistical analysis it is clear that the assumption of independence is violated. However if a statistician was simply presented with the 200 blood pressure measurements, and not told the way in which the data had had been obtained, it would be extremely hard for them to determine that the assumption of independence had been violated.

Notice that assumptions 3) and 4) both concern the **true** residuals, defined in terms of deviations from the model defined by population parameters (such as α and β). However true residuals can never be observed in practice and so to investigate assumptions we have to use the **observed** residuals (replacing population parameters such as α and β by their estimates $\hat{\alpha}$ and $\hat{\beta}$). In fact, as will be explained in 6.5.1, observed residuals are neither independent nor do they have constant variance. However, in most settings the departures from independence and homoscedasticity are so small that we can proceed as if the observed residuals were the true residuals in investigating assumptions 3) and 4).

6.4 Using plots to investigate assumptions

It is recommended that data is always explored using a number of simple plots. In the following sections we explore the most useful plots for both simple and multiple linear regression models.

6.4.1 Scatter plots of the dependent variable against predictor variables

For simple linear regression models a simple plot of the dependent variable against the predictor can usually make serious violations of assumptions apparent. Such plots are particularly good for identifying heteroscedasticity, non-linearity and outliers (points which lie atypically far from the regression line (figure 1).

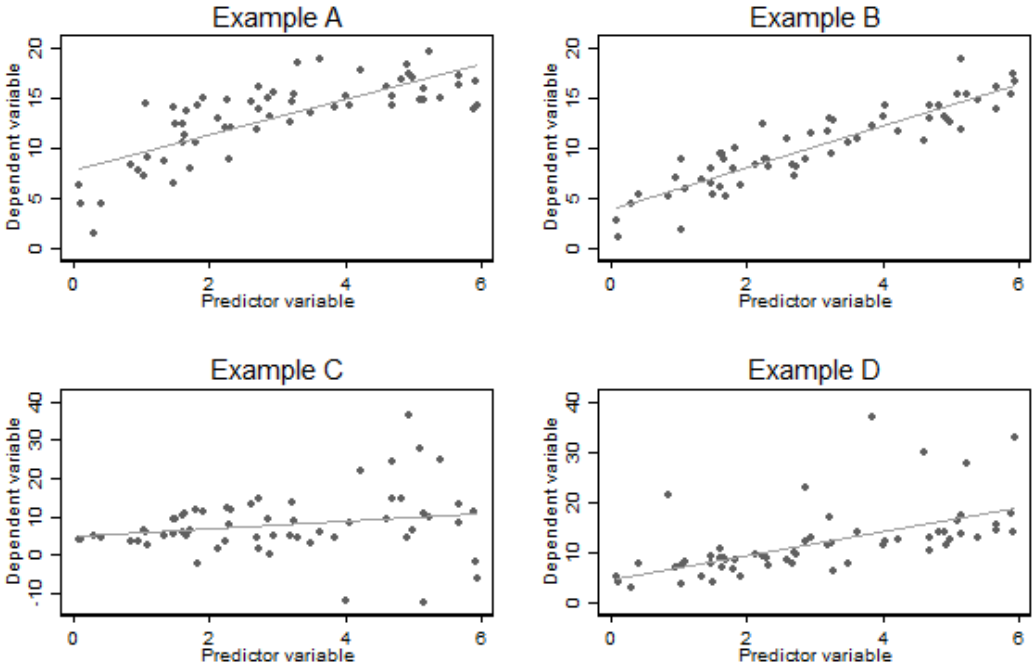


Figure 1: Illustration the usefulness of scatter plots of the dependent variable against the predictor variable in simple linear regression.

Exercise: Which of the graphs in figure 1 provide visual evidence of i) non-linearity, ii) heteroscedasticity, iii) outliers?

In figure 1 the fitted regression lines have been superimposed on the data points. This can be helpful, but is not always so. There is a danger that the inclusion of the line falsely encourages the visual impression of a good fit. It can be a good idea to plot the data both with and without the line of best fit.

For a multiple regression model the linearity assumption in multiple regression requires that the relationship between dependent variable and each predictor is linear conditional on the other predictors in the model. So, there is no requirement that the relationship between the dependent variable and each individual predictor is a linear one when the other predictor variables are ignored. Particularly when there are strong dependencies between predictor variables it is quite possible that relationships between the dependent variable and one or more of the predictor variables are markedly non-linear when other variables are ignored, but are linear conditional on the other variables (*i.e.* the multiple regression model is valid). This means that definitive assessment of the fit of a multiple regression model cannot be inferred from a series of scatter plots relating the dependent variable to each predictor variable. Such plots do have utility in detecting points with extreme values, but for multivariable models they are generally less useful than the residual plots considered in the next section.

6.4.2 Plots of residuals against predictor variables and against fitted values

If the assumptions of the linear regression model hold then the true residuals should be normally distributed with constant variance, and consequently the observed residuals approximately so. Therefore, the homoscedasticity assumption can be investigated by examining plots of the observed residuals against the fitted values. Figure 2 shows residuals plotted against fitted values for the examples in Figure 1 and we can see that the heteroscedasticity for example C is again apparent.

Plots of the observed residuals against the fitted values also allow us to investigate the assumption of linearity. If there is a non-linear relationship present then, as shown in Example A in Figure 2, residuals will not be equally distributed above and below zero across the range of fitted values.

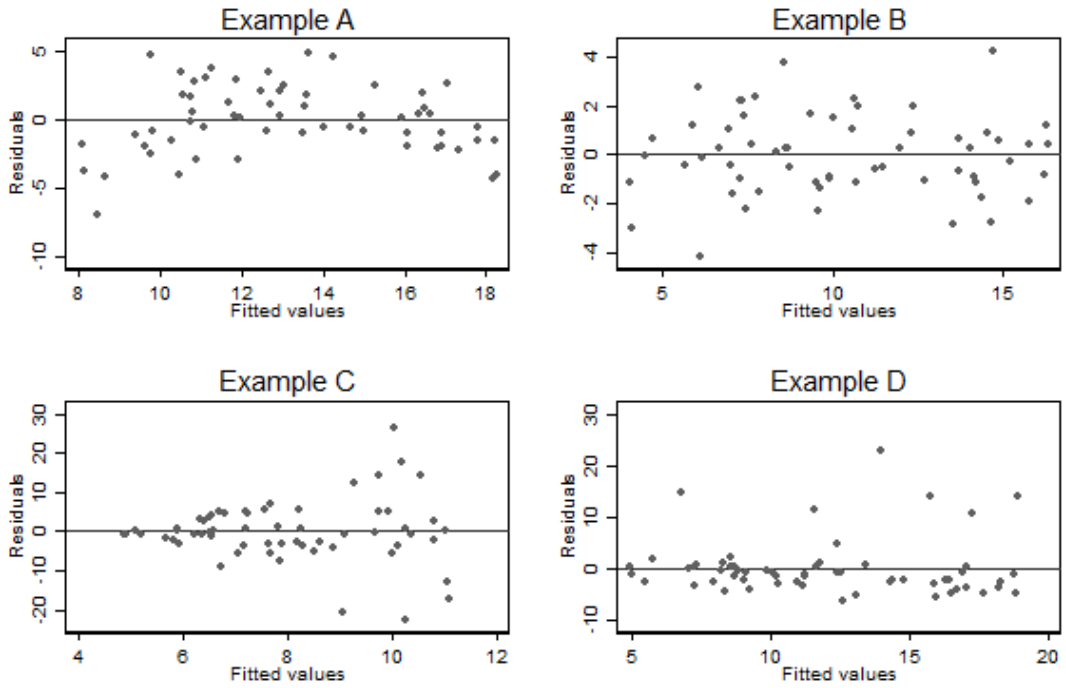


Figure 2: Plots of residuals against fitted values for the examples in Figure 1

For a simple linear regression model the fitted values are a linear transformation of the single predictor variable and so plots of residuals against fitted values are essentially equivalent to plots of residuals against the explanatory variable. This means that for simple linear regression models plots of residuals against fitted values are rarely more revealing than scatter plots of the dependent against the predictor variable. The departures from assumptions evident in Figure 1 are those seen in Figure 2 and vice-versa.

However, for multiple regression models plots of residuals against fitted values can be extremely useful in detecting violations of the homoscedasticity and linearity assumptions. It can also be useful to plot residuals against each predictor variable, as a further check whether there is linear relationship between the dependent variable (Y) and each of the predictor variables (conditional on the other predictor variables in the model). If there are only a small number of predictor variables in the model than these plots can be done for all variables.

However, if the model is very complex it may be judged sufficient to only plot residuals against fitted values and residuals against the most important explanatory variables.

In Stata the **`rvfplot`** command can be used to construct a plot of residuals against fitted values. This command can be used following the use of a **`regress`** command. The resultant graph always relates to the most recently fitted model.

6.4.3 Normal plots of residuals

Normal plots were introduced in Analytical Techniques 6. They provide the best means of visually detecting departures from normality. Normal plots of residuals for the examples in Figure 1 are shown in Figure 3. The results illustrate that normal plots provide an effective means of detecting outliers (and, indeed, are useful for detecting skewness and kurtosis), but are not effective at detecting either heteroscedasticity or non-linearity (since neither of these will necessary result in residuals that markedly depart from a normal distribution).

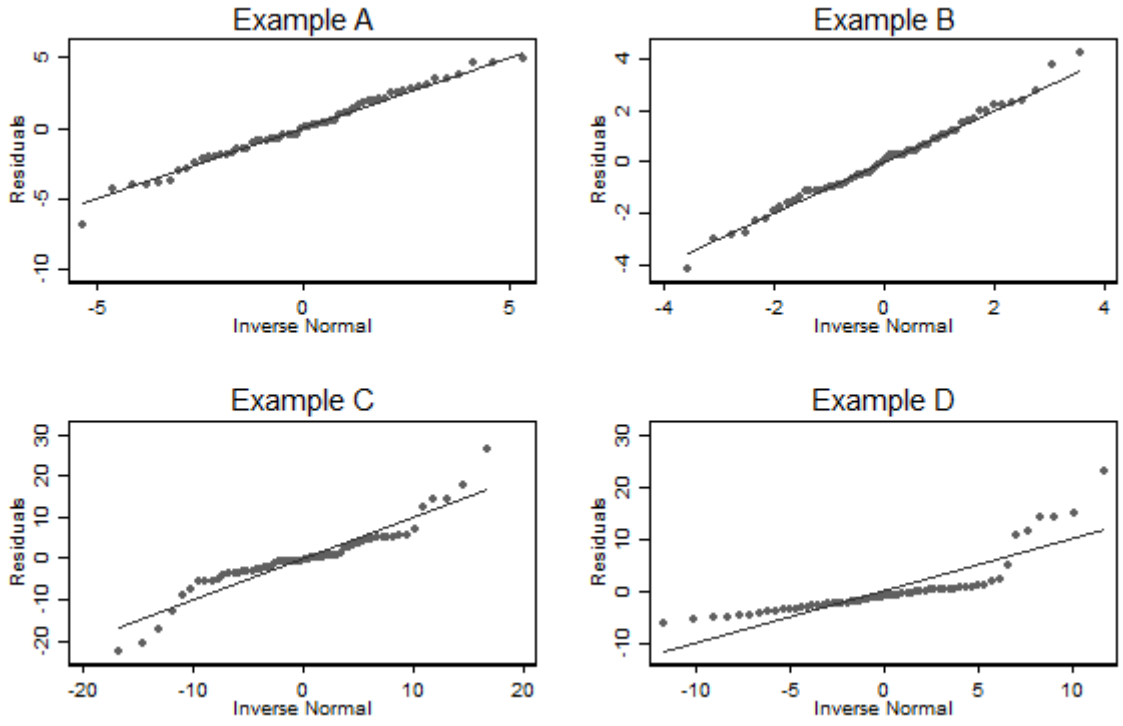


Figure 3: Normal plots of residuals for the examples in Figure 1

As mentioned above and discussed in 6.6.1 below observed residuals do not have constant variance even when true residuals do. Some authors therefore suggest replacing the normal and scatter plots of residuals discussed in the above sections by equivalent plots involving **standardised** residuals (which do have constant variance) (see 6.6.1). In practice the differences between the two approaches are minor and some statisticians prefer to work with the observed residuals since these have the same units as the dependent variable.

6.4.4 Example of residual plots for analysis of cross-sectional study of Gambian children

Figure 4 shows residual plots for the multiple regression model (fitted in Regression 4) relating weight to age and length for the cross-sectional data from the Gambia. None of the three scatter plots are strongly suggestive of either non-linearity or heteroscedasticity. There are also no clear outliers present. Figure 4 also shows a normal plot of residuals. There is some departure from normality, but the departure is not substantial and most statisticians would consider the departures not to be sufficiently large to cast material doubt on the robustness of the inferences made from the model fitted here.

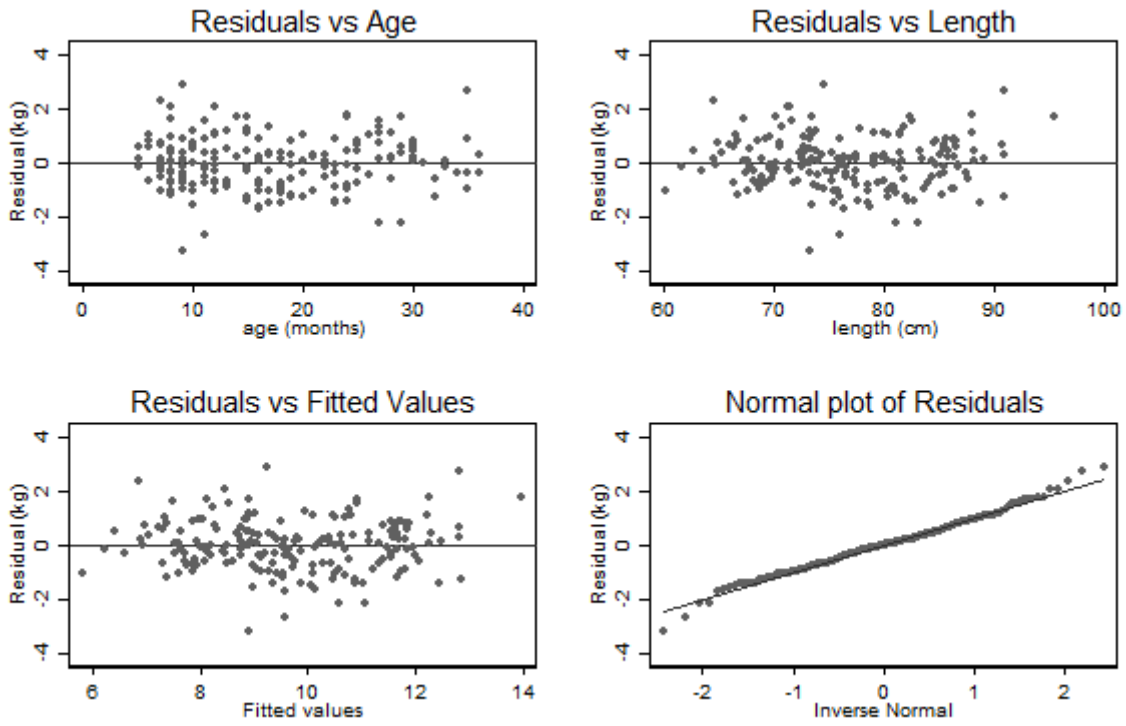


Figure 3: Residual plots for the linear regression relating a child's weight to their age and length for the data from the Gambian cross-sectional study.

Exercise: Is the departure from normality evident in the normal Plot in Figure 3 characteristic of skewness or kurtosis (and if kurtosis, of what type?)?

6.5 Statistical tests of assumptions

It might be anticipated that the assumptions of the linear regression model can best be investigated using formal hypothesis tests. Indeed there exist a number of statistical tests for normality, including the Kolmogorov-Smirnov test and the Shapiro-Wilk test (see Armitage, Berry and Matthews for details) which could be used to test the normality of residuals (or of standardised residuals). Further there exist statistical tests for heteroscedasticity of residuals (for example the **`hettest`** command in Stata (when used after **`regress`**) reports the results of a test of heteroscedasticity due to Breusch-Pagan/Cook-Weisberg).

However (as explained with reference to tests of normality in Analytical Techniques 6) these tests suffer from the drawback that they tend to only have statistical power to detect model violations when datasets are large; and when datasets are large the central limit theorem means that the consequences of these violations of assumptions are less important than in small datasets. With large datasets, tests of normality and heteroscedasticity can often be statistically significant, but the impact of these violations may be practically unimportant. For this reason the tests are considered by many statisticians to be of limited practical use and so details of these procedures will not be given here.

The assumption of linearity can often be usefully tested by fitting a quadratic regression model. This technique is explained in the next section.

6.5.1 The quadratic regression model and its use in testing for non-linearity

The quadratic regression is a multiple regression model with two predictor variables where the second predictor is the square of the first predictor. Algebraically this is:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, where\ \varepsilon_i \sim NID(0, \sigma^2) \qquad (1)$$

Despite the fact that one of the predictors is the square of the other this is still a linear regression model because the expectation of the dependent variable is assumed to be a linear function of the parameters.

The quadratic regression model describes a non-linear relationship between *Y* and *X*. The following Stata output shows the results of fitting such a model, in the REPAIR randomised controlled clinical trial in living-donor renal transplant patients. This model relates a recipient's glomerular filtration rate (GFR) at 12 months after transplant (`gfr_iohexol`) to the duration of ischaemia of the transplanted kidney during the transplant surgery (`time`).

```
. gen time2=time*time

. regress gfr_iohexol time time2

      Source |       SS           df       MS      Number of obs   =       228
-------------+----------------------------------   F(2, 225)       =      9.90
       Model |  4759.10085         2  2379.55043   Prob > F        =    0.0001
    Residual |  54056.6104       225  240.251602   R-squared       =    0.0809
-------------+----------------------------------   Adj R-squared   =    0.0727
       Total |  58815.7112       227   259.10005   Root MSE        =      15.5

------------------------------------------------------------------------------
 gfr_iohexol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.1682453   .0491752    -3.42   0.001    -.2651483   -.0713424
       time2 |      .0004   .0001638     2.44   0.015     .0000772    .0007228
       _cons |    72.8222    3.40221    21.40   0.000     66.11793    79.52647
------------------------------------------------------------------------------
```

Interpreting $\hat{\beta}_1$ and $\hat{\beta}_2$ in a quadratic regression model is not as straightforward as in most linear models with two predictor variables. The reason for this is that it is not possible to change X^2 by 1 unit whilst holding *X* constant (although it is possible to change *X* by 1 unit (from -0.5 to 0.5) whilst holding X^2 constant). One useful way of interpreting the coefficients is to consider the slope of a tangent to the regression line at a particular value of *X* (*X*=*x*), which is $\hat{\beta}_1 + 2\hat{\beta}_2 x$

Exercise: Show that $\hat{\beta}_1 + 2\hat{\beta}_2 x$ is the slope of a tangent to the quadratic regression line $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$.

This makes it clear that the linear term (β_1) in the regression model describes the 'instantaneous' slope of the relationship between *Y* and *X* when *X*=0 (since the tangent at this point is equal to β_1). The quadratic term (β_2) describes the way in which this 'instantaneous' slope changes with increasing *X*. From the model in the REPAIR example, there is some evidence of curvature on the basis that p=0.015 for the quadratic term in the model.

When interpreting the results of fitting a quadratic regression model it is helpful to plot the fitted values as illustrated in Figure 5. The estimated quadratic term in the model is positive, showing that the fitted 'instantaneous' slope is becoming more positive with increasing *X*. It starts with a downward slope with short ischaemia times, since $\hat{\beta}_1 = -0.168$, reaching a minimum at 210 minutes of ischaemia time before becoming increasingly positive above 210 minutes.

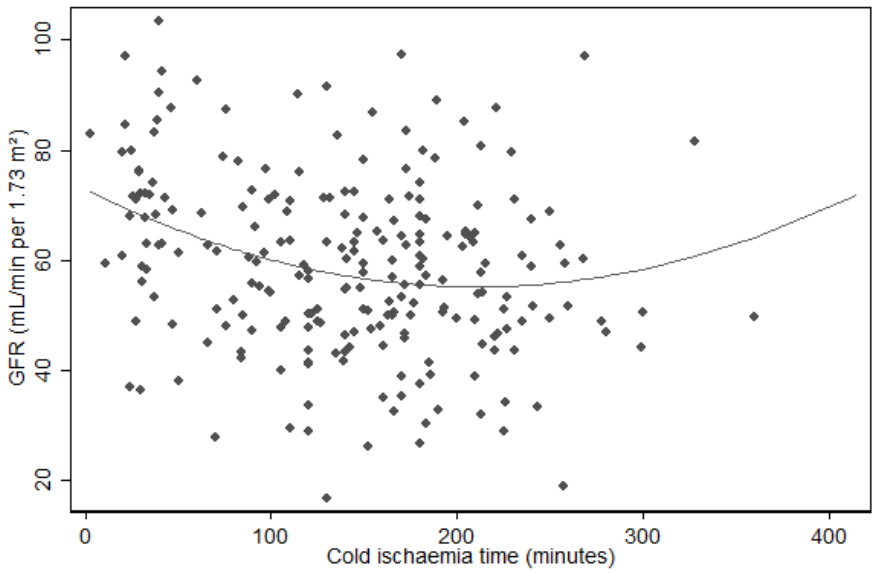


Figure 5: Scatter plot showing fitted values from a quadratic regression model relating a kidney transplant recipient's glomerular filtration rate to duration of ischaemia of the transplanted kidney.

Quadratic terms can be added to regression models with more than one predictor in order to investigate curvature in the relationship between the dependent variable and one predictor, whilst controlling for others. The Stata output below shows the effect of fitting such a model to the cross-sectional data from the Gambia.

```
. regress wt len age

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         len |   .2371289   .0195165    12.15   0.000     .1986281    .2756296
         age |  -.0112598   .0167512    -0.67   0.502    -.0443055    .0217859
       _cons |  -8.351244   1.259968    -6.63   0.000    -10.83682   -5.865667
------------------------------------------------------------------------------
. regress wt len age age2
```

```
------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         len |   .2514685   .0205039    12.26   0.000     .2110183     .2919186
         age |  -.1110198    .050205    -2.21   0.028    -.2100643    -.0119752
        age2 |   .0023351   .0011091     2.11   0.037     .0001471      .004523
       _cons |  -8.591843   1.253778    -6.85   0.000    -11.0653    -6.118391
------------------------------------------------------------------------------
```

Exercise: Interpret the results of fitting the above quadratic regression model.

6.5.2 Modelling non-linear associations

Quadratic regression models have limitations as realistic descriptions of relationships between variables. Quadratic functions either increase to a maximum and then decline: or fall to a minimum and then increase. Further the behaviour of a quadratic is symmetric about the 'turning point'. Such relationships are implausible in many medical applications. Although the fitted 'turning point' can be outside the range of observed values of X, it often is not, leading to fitted relationships that are a-priori unrealistic. For example, the model fitted for the REPAIR data suggests that kidney function increases as duration of ischaemia increases after 210 minutes, which seems implausible. Therefore, if a quadratic regression model provides evidence against a linear relationship and the curvature appears substantial this will often motivate fitting a more realistic model to better describe the relationship.

Further power terms may be included to fit a polynomial regression model. The quadratic regression model is the simplest of these types of model. The next simplest is the cubic model in which a cubic term is added to the model described by equation (1). An even more flexible approach is to use a piecewise polynomial model, which allows for a different polynomial function in different ranges of the observed values of X, defined according to specified "knots". These kind of models can be written as a multiple regression model with additional predictor variables that are a transformation of the first predictor. One approach is a restricted cubic spline model, where the fitted mean function is (i) linear before the first knot and after the last knot, (ii) a piecewise cubic polynomial function between the remaining knots, and (iii) has a smooth function by constraining the first and second derivatives of adjacent functions to agree when they meet at the knot point. The `mkspline` command in Stata can be used to create the necessary predictor variables to fit a restricted cubic spline for a given predictor variable. An example of a restricted cubic spline model is illustrated in Figure 6 for the REPAIR data.

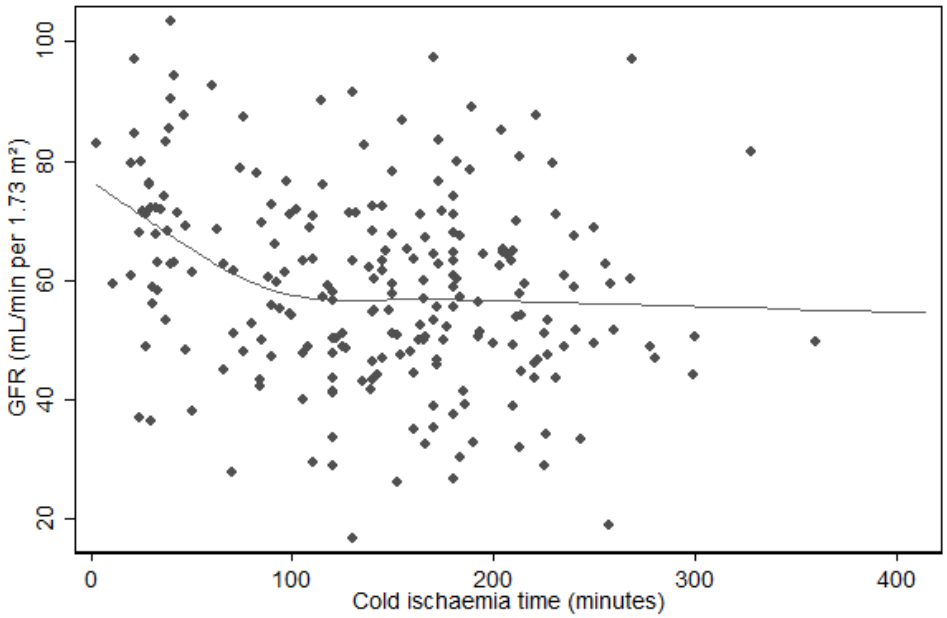


Figure 6: Scatter plot showing fitted values from a restricted cubic spline regression model relating a kidney transplant recipient's glomerular filtration rate to duration of ischaemia of the transplanted kidney, with knots at 30, 90, 150 and 210 minutes

6.6 Outliers, Leverage, Influence and Cook's Distance

When fitting a linear regression model it is sensible to consider the impact that individual observations can have on the fitted results. In a loose sense it is important to identify points that have extreme values of the dependent variable (**outliers**), points that have extreme values of the predictor variables (points with large **leverage**) and points which exert particular **influence** on the parameters of the fitted regression model (measured by **Cook's Distance**).

6.6.1 Outliers and standardised residuals

Outliers are points that lie 'a long way' from the fitted value under the regression model. It is important to identify such points for a number of reasons. First outliers can be mistakes that warrant further investigation (for example checking of source data). Second outliers can have a big impact on standard errors. Third outliers can (but do not always) have an impact on the parameter estimates.

The distance between an observed value and the fitted value is the observed residual. Even though one of the assumptions of the linear regression model is that the true residuals are independent and have constant variance, observed residuals are not independent and have non-constant variance. To understand why observed residuals are not independent consider fitting a regression model to data in two groups, where one of the two groups contains only two observations. The fitted value in this group will be the arithmetic mean of the two measures and so the two observed residuals will be equal in magnitude and have opposite directions (*i.e.* they will be perfectly negatively correlated). In the general situation the same effect is observed *i.e.* observed residuals for observations with similar values of the predictor variables will be negatively correlated with each other because the fitted values from the regression model will tend to pass close to the 'local' mean of the observations.

The observed residuals do not have constant variance because points that are extreme in terms of their predictor variables will tend to 'pull' the fitted values towards themselves. Again consider fitting a regression model to two groups, but this time where one of the groups contains only a single observation. The residual variance for this observation is zero because the fitted value will be the observed value for this observation. In the simple regression model the variance of residuals is given by:

$$V(\hat{\varepsilon}_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SS_{xx}}\right] \qquad (2)$$

From equation (2) it is clear that the residual variance decreases with increasing distance from the mean of the predictor variable. Additionally, the variance approaches σ^2 as n increases.

In the multiple regression notation introduced in Regression 4 and 5, the variance-covariance matrix of the residuals takes the following form.

$$V(\hat{\varepsilon}_i) = \sigma^2[I - P] \; where \; P = X(X'X)^{-1}X' \qquad (3)$$

The diagonal elements of this matrix are the variances, with the off-diagonal elements representing covariances. As mentioned previously, since ***P*** is not, in general, a diagonal matrix the observed residuals are not independent.

Standardised residuals (r_i) are residuals divided by their estimated standard errors (obtained by replacing σ^2 with its estimated value ($\hat{\sigma}^2$) in equations (2) and (3)). Standardised residuals have unit variance and quantify the extent to which a point is an **outlier**. In assessing whether or not points are outliers it is useful to note that approximately 95% of observations should have standardised residuals that are less than 2 in magnitude and that approximately 99.7% of points should have standardised residuals that are less than 3 in magnitude. Figure 7 shows a dataset with two outlying values. The standardised residuals for the two outliers are indicated on the figure. It is apparent that in this example these two outliers have had a greater impact on the fitted intercept (most of the points lie below the fitted regression line) than on the slope.

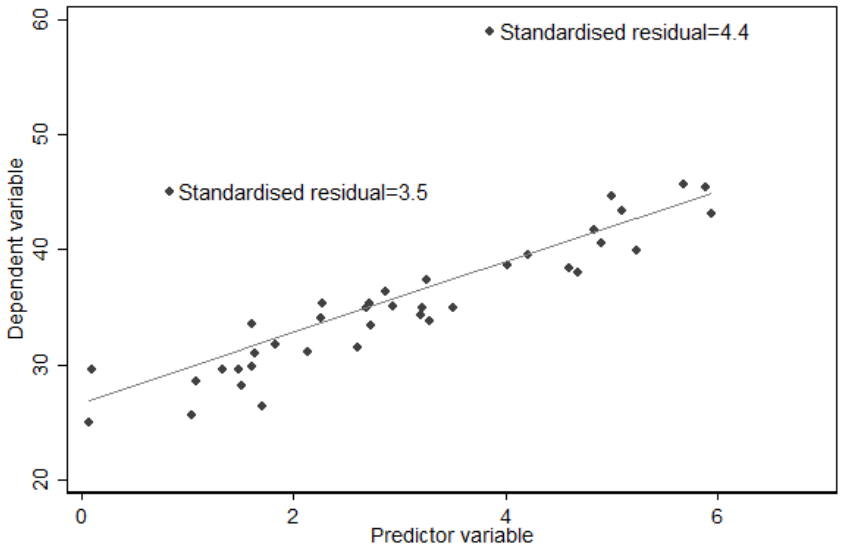


Figure 7: Scatter plot showing fitted values for a dataset with two outlying values.

An alternative to the standardised residual is the **studentised residual**. Studentised residuals are residuals divided by their estimated standard error, with σ^2 separately estimated for each observation. In each case σ^2 is estimated from a regression model omitting the point in question (to avoid the inclusion of the point in question falsely inflating the estimated residual variance). For the two points highlighted in Figure 7 the studentised residuals are 4.2 and 6.3 respectively.

6.6.2 Leverage

Observations with extreme values of the predictor variable(s) have the potential to have a strong impact on parameter estimates. The extent to which the predictor variables for a particular observation are extreme is measured by its **leverage**. In a simple linear regression model the leverage of the *i*th observation is defined as follows:

$$l_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_{xx}} \qquad (4)$$

Leverage in this simple situation is a monotonic function of the distance between the value of the predictor variable and the mean. In multiple regression the leverage is defined as follows:

$l_i = \boldsymbol{P}_{ii}$ the *i*th diagonal element of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (5)

Leverage takes values between 1/*n* and unity. The larger the value the greater the potential influence over the fitted values. If the value is equal to 1 the influence is so strong that the fitted value will be equal to the observed value.

In multiple regression leverage measures 'distance' from the centre of the joint distribution of the predictor variables, but with distance scaled by the directional degree of dispersion. This is illustrated in Figure 8 which shows a scatter plot of two predictor variables. In a multiple regression model including both of these predictor variables there are two points with particularly large leverage (the leverage for the point with the third highest value is 0.10).

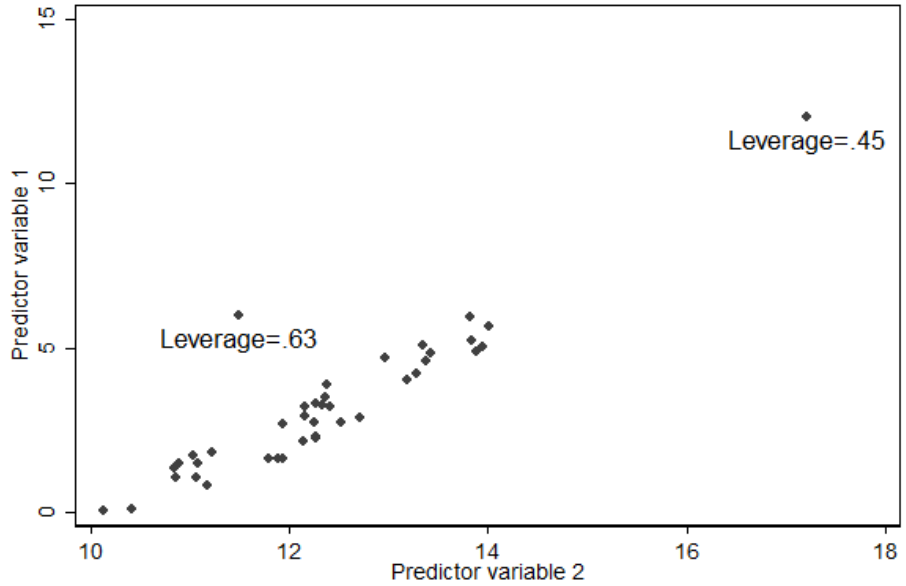


Figure 8: Scatter plot illustrating points with large leverage in a multiple regression model that includes two predictor variables.

Notice that the point with largest leverage in Figure 8 would not have particularly high leverage in either simple linear regression models including only one of the two predictor variables. Further this point would not be readily identified in plots of each of the predictor variables against the dependent variable. The value of measures such as the leverage is greatest in complex multiple regression models where it can be difficult to identify points with an atypical combination of predictor variables using more simple graphical techniques.

6.6.3 Influence and Cook's Distance

Points with large leverage are best thought of as being 'potentially influential'. The values that their predictor variables take mean that they can have a large impact on parameter estimates and fitted values. However they need not necessarily be influential in practice: it is possible that removal of a point with large leverage from a dataset will not materially alter the fitted values. In contrast **Cook's distance** measures the actual **influence** that exclusion of a point has on the fitted values.

For a model with p predictor variables (with estimated residual variance $\hat{\sigma}^2$) the Cook's distance for the ith observation (D_i) is constructed by refitting the model excluding this observation and obtaining new fitted values ($\hat{y}_{j(i)}$) for all n observations (including the omitted one). D_i is then as follows:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{(p+1)\hat{\sigma}^2} \qquad (6)$$

It can be shown that Cook's distance is a function of the standardised residual (r_i) and the leverage (l_i). Specifically

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{(p+1)\hat{\sigma}^2} = \frac{r_i^2 l_i}{(p+1)(1-l_i)} \qquad (7)$$

It is clear from this formula that the points which have most influence on the fitted values are those which combine large leverage (l_i) with a large standardised residual (r_i).

Different authors give different recommendations as to how large D_i should be in order to warrant further investigation. However one sensible approach is to identify a group of points with values of D_i that are substantially larger than the others, refit the model omitting these points, and assess the extent to which inferences from the model are materially altered.

It can often be informative to display Cook's distances, leverages and standardised residuals graphically. In particular plots of i) Cook's distances against fitted values and ii) leverages against squared standardised residuals are sometimes used.

6.6.4 Stata commands

Stata's **predict** command can be used to obtain standardised residuals, leverages, Cook's Distances and other diagnostic measures. The basic syntax (used after the **regress** command) is:

```
predict <newvar>, < option >
```

where **newvar** is the name of the newly derived variable and **option** determines the type of statistic that is required. For example:

```
predict c, cooksd        creates a variable c that contains Cook's distances.
predict l, leverage      creates a variable l that contains leverages.
predict r, rstan         creates a variable r that contains standardised residuals.
```

6.7 Dealing with violations of assumptions

In this session we have focussed on the identification of the violation of model assumptions through the use of diagnostic plots, statistical tests and the identification of outlying values. We have also explained how to identify points that have the strongest influence on the results from a regression model. This identification of issues of potential concern is only the first, and arguably the easiest, aspect of an exploration of the robustness of the results of fitting a linear regression model. In this section we briefly describe some approaches that can be adopted, motivated by findings arising from use of the techniques introduced above.

6.7.1 Data checking

It is obviously important that mistakes in data be eliminated as far as possible. In practice ensuring that a large dataset is 100% error free may be impossible. Observations with large standardised residuals can potentially arise through data entry or coding errors and so usually the first step is to check such values with the data provider or original source data if available. Furthermore observations with large Cook's distances are those which have most impact on the results and so it is always a good idea to double-check that such values are correct.

6.7.2 Sensitivity analysis

Where there is a particularly influential observation, or group of observations, it is a good idea to repeat the analysis omitting that observation (or those observations) and to discuss the extent to which the results are altered.

6.7.3 Transformations

Sometimes it can be useful to transform either the dependent variable and/or one or more of the predictor variables. There can be a number of motivations for this

i) As mentioned in **section 6.5.2** this may be done to convert a non-linear relationship into a linear model. For example:

$$y_i = \alpha(x_i)^{\beta} \Rightarrow \log(y_i) = \log(\alpha) + \beta \log(x_i).$$

The advantage of this is that non-linear models are typically more complicated to fit than linear ones.

ii) Sometimes a transformation may stabilise the variance of residuals.

iii) Sometimes a transformation can improve the normality of residuals.

Transformations are discussed in some detail in Analytical Techniques 6. It is important to realise that in a linear regression model it is only the residuals that are assumed to be normally distributed. There is never any requirement that the predictor variables in a model be normally distributed. Furthermore the dependent variable need not be normally distributed, although in situations where R^2 is small approximate normality of residuals is usually synonymous with approximate normality of the dependent variable.

6.7.4 Non-linear models

Sometimes the techniques introduced in this session suggest that a non-linear model is more appropriate than a linear one. Non-linear models are models in which the expectation of the dependent variable is not a simple linear combination of the parameters in the model. Such models are beyond the scope of this module, but are introduced later in the course.

6.7.5 Robust methods

As mentioned in Analytical techniques 6 there are a number of relatively computer intensive methods (*e.g.* Bootstrapping, 'Sandwich' estimators of variance) of calculating standard errors that relax assumptions concerning normality of residuals and/or homogeneity of residual variance and/or independence of residuals. See Robust Methods for details.

Many of the issues raised in this session will be revisited in the Generalised Linear Models module in term 2, where strategies for building useful and reliable models will be considered in some detail.

Regression 7: Interactions

7.1 Objectives

By the end of this session students will be able to:

- Explain the motivation for investigating interactions.
- Fit and interpret linear regression models that include interaction terms using a statistical package.

7.2 Introduction

In previous sessions we have discussed how we can use multiple regression to control (or adjust for) the effects of other variables when investigating the relationship between a dependent variable (Y) and a particular predictor variable (X). The aim of such an analysis is to estimate the slope of the association between Y and X whilst holding all the other predictor variables constant.

Such models assume that the slope of the association between Y and X is the same whatever the values of the other predictor variables in the model. For example if we use a multiple regression model to relate children's weight to their age, gender and height then the partial regression coefficient for height represents the effect of a unit increase in height on weight in children of the same age and gender. It is implicitly assumed by the form of the model that the coefficient relating weight to height for is the same for children of all ages and genders, for example that it is the same for two-year old girls as in three-year old boys. Of course this need not necessarily be the case. It could be that the slope of the association between weight and height differs by gender and/or by age.

The term **interaction** is used to describe situations in which the relationship between Y and X differs according to the level of one or more other predictor variables. In linear regression models it is the **slope** of the relationship between Y and X that depends on other factors. In other setting it might be another parameter (such as an odds ratio) that depends upon other factors.

In this session we will consider how to extend linear regression models to include such interactions and how to interpret the results. We will consider interactions between pairs of continuous variables, pairs of categorical variables and pairs of variables one of which is categorical and the other continuous (the simplest case conceptually).

7.3 Linear regression model with an interaction between two predictor variables

7.3.1 General formulation of the model

Suppose we wish to relate a dependent variable (*Y*) to two predictor variables (X_1 and X_2) but wish to allow the slope of the association between *Y* and X_1 to differ according to the value of X_2. To allow for this we fit an interaction model that contains a predictor variable (X_3) that is the product of X_1 and X_2.is as follows.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \text{ with } \varepsilon_i \sim NID(0, \sigma^2) \quad (1)$$

where,

y_i = value of the dependent variable
x_{1i} = value of the first predictor variable
x_{2i} = value of the second predictor variable
$x_{3i} = x_{1i} \times x_{2i}$

To understand the why this model allows the slope of the association between *Y* and X_1 to vary according to X_2, we can consider the form that equation (1) takes for when we fix X_2 to have a particular value (say $X_2 = k$). In this situation the relationship between *Y* and X_1 is as follows.

$$y_i = (\alpha + \beta_2 k) + (\beta_1 + \beta_3 k) x_{1i} + \varepsilon_i \quad (2)$$

For these observations the relationship between *Y* and X_1 is a linear one with both slope and intercept dependent upon *k.* The intercept is $\alpha + \beta_2 k$ and the slope is $\beta_1 + \beta_3 k$.

By allowing the association between *Y* and X_1 to vary according to X_2, we have also allowed the slope for the association between *Y* and X_2 to vary according to X_1. If we look at the form of equation (1) takes for a particular value of X_1 (say $X_1 = l$) we find:

$$y_i = (\alpha + \beta_1 l) + (\beta_2 + \beta_3 l) x_{2i} + \varepsilon_i \quad (3)$$

Again the relationship between *Y* and X_2 is a linear one with both slope and intercept dependent upon *l.*

7.3.2 Interaction between a continuous predictor variable and a binary predictor

The interaction model is particularly easy to interpret when one of the predictor variables (say X_2) is a binary predictor variable taking the values 0 and 1 (*i.e.* a dummy variable). Equation (2) then becomes

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{1i} + \varepsilon_i && \text{when } X_2 = 0 \\ y_i &= (\alpha + \beta_2) + (\beta_1 + \beta_3) x_{1i} + \varepsilon_i && \text{when } X_2 = 1 \end{aligned} \quad (4)$$

For both values of X_2 the relationship between *Y* and X_1 is a linear one with different slopes and intercepts in the two groups. The interpretation of each of the parameters is as follows.

α is the intercept when $X_2 = 0$.

$\alpha + \beta_2$ is the intercept when $X_2 = 1$ and hence β_2 is the difference in intercepts between the two groups defined by X_2.

β_1 is the slope when $X_2 = 0$.

$\beta_1 + \beta_3$ is the slope when $X_2 = 1$ and hence β_3 is the difference in slopes between the two groups defined by X_2.

As an example consider an ecological study relating mean systolic blood pressure blood pressure to mean twenty four hour salt intake in twenty four communities, twelve of which were characterised as low income countries and twelve as high income countries. Mean blood pressures and salt intakes were estimated from studies involving large numbers of people in each of these communities. Figure 1 displays the data.

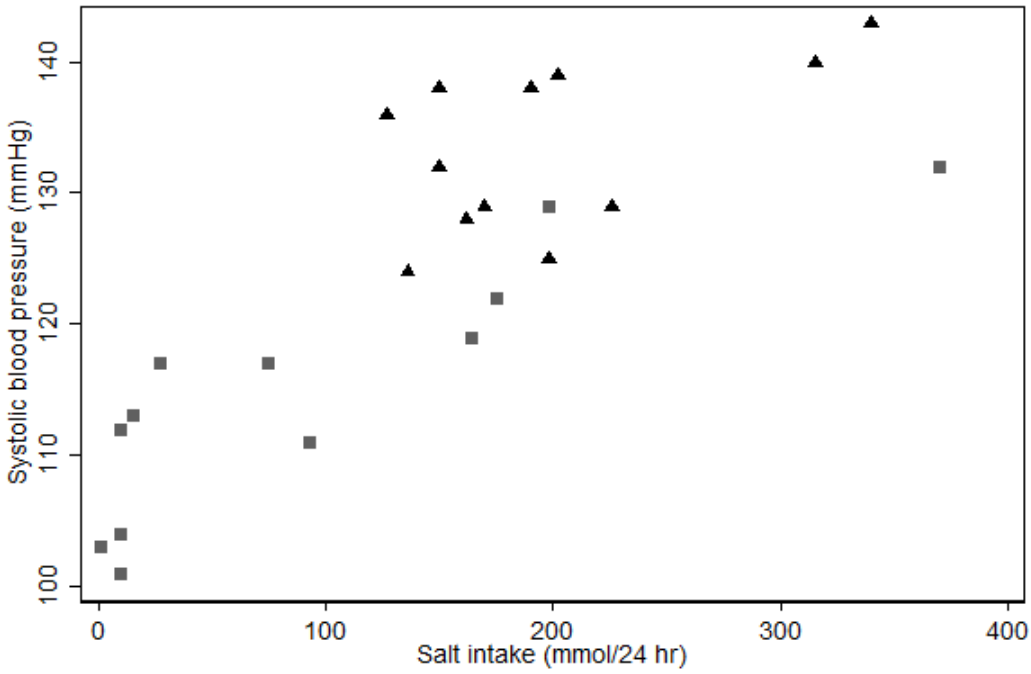


Figure 1: Data from an ecological study relating mean blood pressure to mean salt intake and economic status in twelve high income communities (triangles) and twelve low income communities (squares).

The Stata output on the next page shows the effect of fitting simple and multiple regression models with systolic blood pressure (mmHg) as the dependent variable and salt intake (mmol/24 hr) and economic status (0 = low income, 1 = high income) as predictor variables.

Exercise: Explain the differences between the crude and adjusted regression coefficients for salt intake and economic status. Do you think this is this indicative of confounding?

```
. regress bp econ
```

Model A

```
------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        econ |   18.41667   3.363031     5.48   0.000     11.44217    25.39117
       _cons |         115   2.378022    48.36   0.000     110.0683    119.9317
------------------------------------------------------------------------------
```

```
. regress bp salt
```

Model B

```
------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salt |   .0963668   .0146867     6.56   0.000     .0659086    .1268251
       _cons |   110.0986   2.623689    41.96   0.000     104.6574    115.5398
------------------------------------------------------------------------------
```

```
. corr salt econ
(obs=24)
             |     salt     econ
-------------+------------------
        salt |   1.0000
        econ |   0.4958   1.0000
```

```
. regress bp salt econ
```

Model C

```
------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salt |   .0686311   .0122734     5.59   0.000     .0431071    .0941551
        econ |   11.45061   2.512426     4.56   0.000     6.225733    16.67549
       _cons |   108.4343   1.938773    55.93   0.000     104.4024    112.4662
------------------------------------------------------------------------------
```

Fitted values from model C are illustrated in Figure 2. The fitted values form two straight lines with common slope 0.069 mmHg/(mmol/24 hr) and intercepts 108.4 mmHg and 119.9 mmHg (119.89 = 108.43 + 11.45) respectively.

This model permits estimation of the effect of salt intake on blood pressure adjusting for the effect of economic status. It also permits estimation of the effect of economic status on blood pressure adjusting for the effect of salt intake.

Model C is an example of an Analysis of Covariance (ANCOVA) model (see Regression 4), also sometimes termed a ‘parallel lines’ regression model. It permits adjustment of the effect of one predictor variable for the effects of others, but forces the effect of a unit change in each predictor variable to be constant, whatever the level of the other predictor variables (*i.e.* the lines in figure 2 are constrained to be parallel).

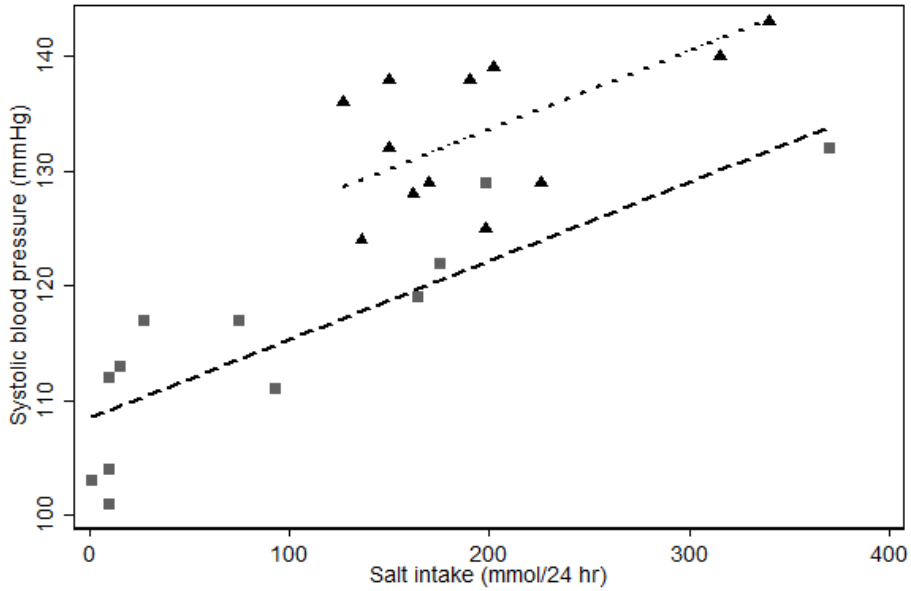


Figure 2: Data and fitted values from a parallel lines regression model for the ecological study. Fitted values for the twelve high income communities (triangles) shown as a dotted line. Fitted values for the twelve low income communities (squares) shown as a dashed line.

To fit an interaction term a new predictor variable, equal to the product of the existing predictor variables must be added to model C. This can be achieved as follows in Stata.

```
. gen inter=salt*econ

. list bp salt econ inter

     +---------------------------+
     |  bp   salt   econ   inter |
     |---------------------------|
  1. | 103      1      0       0 |
  2. | 101     10      0       0 |
   .     .       .      .       .
   .     .       .      .       .
 12. | 132    370      0       0 |
 13. | 136    127      1     127 |
 14. | 124    136      1     136 |
   .     .       .      .       .
   .     .       .      .       .
 24. | 143    340      1     340 |
     +---------------------------+

. regress bp salt econ inter      Model D

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salt |    .075024   .0144064     5.21   0.000     .0449729    .1050752
        econ |   15.55493    5.39218     2.88   0.009      4.30704    26.80282
       inter |  -.0241075   .0279757    -0.86   0.399    -.0824638    .0342487
       _cons |   107.8227    2.07586    51.94   0.000     103.4925    112.1529
------------------------------------------------------------------------------
```

Fitted values from model D are illustrated in Figure 3. The intercept and slope among the low income communities are 107.8 mmHg and 0.075 mmHg/(mmol/24 hr) respectively. The intercept and slope among the high income communities are 124.1 mmHg (123.4= 107.8 + 15.6) and 0.051 mmHg/(mmol/24 hr) (0.051 = 0.075 + -0.024) respectively. The interaction term has p=0.399 and so we do not have evidence that the slopes are different.

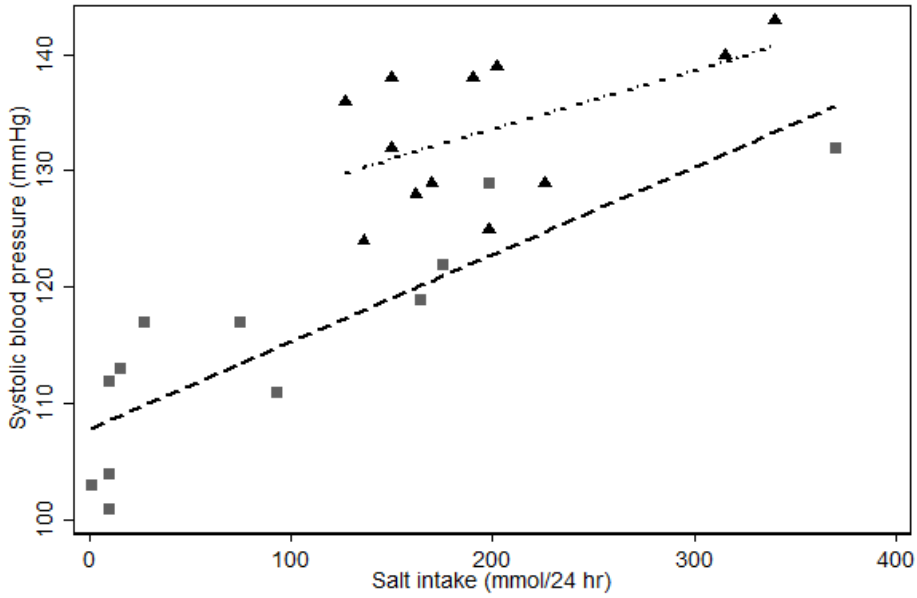


Figure 3: Data and fitted values from an interaction model for the ecological study. Fitted values for the twelve high income communities (triangles) shown as a dotted line. Fitted values for the twelve low income communities (squares) shown as a dashed line.

Note that the same fitted values as in model D could have been obtained by fitting separate linear regression models relating blood pressure to salt intake in both economic status groups. The main advantage of fitting model D is that it provides a test of the null hypothesis that the slopes are the same. Another difference is that the interaction model assumes that the residual variance in the two groups is the same, whereas fitting separate models allows different variances in the two groups. This means that confidence intervals and hypothesis tests derived using the two approaches will differ somewhat.

In using Stata to analyse the above data we created the interaction variable to aid understanding. In fact for categorical predictor variables Stata can do this automatically through the use of the **`i.`** and **`c.`** prefixes in combination with **`#`** or **`##`**, as shown here.

```
. regress  bp salt i.econ c.salt#i.econ    (or. regress  bp c.salt##i.econ)

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salt |    .075024   .0144064     5.21   0.000     .0449729    .1050752
     1.econ |   15.55493    5.39218     2.88   0.009      4.30704    26.80282
            |
 econ#c.salt |
          1  |  -.0241075   .0279757    -0.86   0.399    -.0824638    .0342487
            |
       _cons |   107.8227    2.07586    51.94   0.000     103.4925    112.1529
------------------------------------------------------------------------------
```

The **c.** prefix tells Stata that a variable is a continuous variable that should not be converted to dummy variables before inclusion in the model. Here since **econ** is already a variable that takes the values 0 and 1 it is strictly unnecessary to tell Stata that it is a categorical variable with **i.** , but the syntax above illustrates the general way to deal with categorical predictor variables. The shortened **##** notation tells Stata to include the individual predictor variables as well as the interaction term.

7.3.3 Interaction between two binary predictor variables

When both of the predictor variables are binary predictor variables taking the values 0 and 1 equation (2) can be written as

$$\begin{aligned} y_i &= \alpha + \varepsilon_i && \text{when } X_1 = 0 \text{ and } X_2 = 0 \\ y_i &= \alpha + \beta_1 + \varepsilon_i && \text{when } X_1 = 1 \text{ and } X_2 = 0 \\ y_i &= \alpha + \beta_2 + \varepsilon_i && \text{when } X_1 = 0 \text{ and } X_2 = 1 \\ y_i &= \alpha + \beta_1 + \beta_2 + \beta_3 + \varepsilon_i && \text{when } X_1 = 1 \text{ and } X_2 = 1 \end{aligned} \qquad (5)$$

If we denote the population mean of Y when $X_1 = i$ and $X_2 = j$ by μ_{ij} the interpretation of each of the parameters is as follows.

α is the mean value of Y when $X_1 = 0$ and $X_2 = 0$ (μ_{00}).

$\alpha + \beta_1$ is the mean value of Y when $X_1 = 1$ and $X_2 = 0$. Hence β_1 is the difference in the mean values of Y between the two groups defined by X_1 when $X_2 = 0$ $(\mu_{10} - \mu_{00})$.

$\alpha + \beta_2$ is the mean value of Y when $X_1 = 0$ and $X_2 = 1$. Hence β_2 is the difference in the mean values of Y between the two groups defined by X_2 when $X_1 = 0$ $(\mu_{01} - \mu_{00})$.

$\alpha + \beta_1 + \beta_2 + \beta_3$ is the mean value of Y when $X_1 = 1$ and $X_2 = 1$. Hence β_3 is the difference in the mean values of Y between the two groups defined by X_2 when $X_1 = 1$ <u>minus</u> the difference in the mean values of Y between the two groups defined by X_2 when $X_1 = 0$ $[(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})]$.

Interpretation of β_3 is symmetric in X_1 and X_2: *i.e.* it is also interpretable as the difference in the mean values of Y between the two groups defined by X_1 when $X_2 = 1$ minus the same difference when $X_2 = 0$ $[(\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})]$.

When X_1 was continuous the interaction term we interpreted the interaction term (β_3) as relating to a <u>difference</u> in <u>slopes</u> of Y on X_1 between two groups defined by the binary variable X_2. Here, because X_1 is itself binary the interaction term is interpretable as relating to a <u>difference</u> in <u>differences</u> of Y by categories of X_1 between the groups defined by X_2.

To illustrate such an interaction model we return to the data analysed in the practical session of Regression 4. In the following Stata analysis **seruvitc** is a participant's serum vitamin C level (μmol/l) and **cigs** and **ctakers** are binary predictors taking the values 1 if the participant is a smoker and a taker of vitamin C supplements (and 0 if not) respectively.

```
. regress seruvitc i.cigs##i.ctakers

------------------------------------------------------------------------------
    seruvitc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.cigs | -9.204762   7.372606    -1.25   0.215    -23.85627     5.446744
   1.ctakers |  24.62465   5.919434     4.16   0.000     12.86102     36.38828
             |
cigs#ctakers |
        1 1  | -26.82465   17.79034    -1.51   0.135    -62.1792     8.529904
             |
       _cons |  49.90476   2.728724    18.29   0.000       44.482    55.32753
------------------------------------------------------------------------------
```

Exercise: Interpret each of the estimated partial regression coefficients in this model. What are the mean values of serum vitamin C in the four groups defined by smoking and supplement taking status?

7.3.4 Interaction between two continuous predictor variables

We have seen that when X_1 is continuous and X_2 is binary the interaction term is interpretable as the difference in slopes (Y on X_1) between two groups defined by the binary variable X_2. When both X_1 and X_2 are continuous the interaction term is interpretable as the difference in slopes (Y on X_1) per unit increase in X_2. To illustrate this we return to the cross-sectional data on children's weight and length in the Gambia. The following Stata analysis relates children's weight to their length and age allowing for an interaction between these two predictor variables.

```
. regress wt c.age##c.len

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age | -.2849002   .1049554    -2.71   0.007    -.4919562   -.0778442
         len |  .1877704   .0268079     7.00   0.000     .1348837    .2406572
             |
 c.age#c.len |  .0033981   .0012872     2.64   0.009     .0008588    .0059374
             |
       _cons | -4.519559   1.909179    -2.37   0.019    -8.285989   -.7531303
------------------------------------------------------------------------------
```

To interpret the output first focus on the relationship between weight and length in children aged 12 months. Amongst such children the fitted relationship is as follows.

$$E(weight|age = 12) = (-4.5196 - 0.2849 \times 12) + (0.1878 + 0.0034 \times 12) \times length$$
$$= -7.9384 + 0.2286 \times length$$

Amongst children aged 13 months the analogous relationship is

$$E(weight|age = 13) = (-4.5196 - 0.2849 \times 13) + (0.1878 + 0.0034 \times 13) \times length$$
$$= -8.2233 + 0.2320 \times length$$

Amongst children aged 14 months the analogous relationship is

$$E(weight|age = 14) = (-4.5196 - 0.2849 \times 14) + (0.1878 + 0.0034 \times 14) \times length$$
$$= -8.5082 + 0.2354 \times length$$

The interpretation is that at each age the relationship between weight and length is linear with the slope of the association increasing by 0.0034kg/cm for each one month increase in age.

Interpretation of the other parameters in the above output is made potentially confusing because they represent extrapolations outside the range of the observed data. For example the constant term in the model represents the fitted mean weight in children aged zero and having zero length. The 'length' regression coefficient represents the effect of a 1cm increase in length amongst children aged zero *etc*. To make these parameters more readily interpretable we can centre the predictor variables (see Regression 2) and refit the model. This is illustrated below for the Gambian cross-sectional data.

```
. sum age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       190    16.97895    8.336798          5         36

. gen age_c=age-r(mean)

. sum len

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         len |       190    76.69737    7.155576       60.1       95.5

. gen len_c=len-r(mean)

. regress wt c.age_c##c.len_c

------------------------------------------------------------------------------
          wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_c |  -.0242746   .0172112    -1.41   0.160    -.058229     .0096797
       len_c |   .2454667   .0194701    12.61   0.000     .207056     .2838773
             |
     c.age_c#|
     c.len_c |   .0033981   .0012872     2.64   0.009     .0008588    .0059374
             |
       _cons |   9.469781   .0950748    99.60   0.000     9.282218    9.657345
------------------------------------------------------------------------------
```

Exercise: Interpret each of the estimated partial regression coefficients in this model.

7.3.5 More complex models

Interaction terms between any pair of predictor variables in a regression model can be included in regression models. For example if we have three predictor variables (*A, B, C*) it is possible to simultaneously investigate interactions between each pair of predictor variables by fitting a model that includes each of the three predictor variables together with three interaction terms (*A*B, B*C, A*C*). It would also be possible to include interactions between all three variables (*A*B*C*) which express the extent to which a two-way interaction (such as *A*B*) depends upon a third factor (*C*).

To carry out hypothesis tests relating to interaction parameters we can use the same framework introduced in earlier sessions. In all of the examples considered above only a single interaction term is included in the model and so we could do a Wald test on this parameter in order to test for an interaction. This would be test of the null hypothesis H_0: $\beta_3 = 0$ against the alternative hypothesis H_0: $\beta_3 \neq 0$. If we have multiple interaction terms we can use a partial *F*-test to compare the more complex model containing the interaction terms to the simpler model without these terms. For example, consider a model investigating the relationship between blood pressure and two predictors: age and quartile of income. The simpler (non-interaction) model would have age and three dummy variables to represent the difference in blood pressure between the four levels of income quartile. The more complex model would include these predictors and also three interaction terms for age multiplied by each dummy variable for income quartile. We could then test for an interaction between age and income quartile by using a partial *F*-test to compare the two models. This is a joint test that any of the three interaction terms are non-zero, so a small p-value would suggest an interaction is present and the slope for age differs between one or more of the levels of income. As previously discussed, we might follow this up with further analysis to determine which particular income groups differed in their slopes for age.

If our model relates a number of predictor variables to a dependent variable there is huge scope for including interaction terms. With (only) five predictors there are ten possible two-way interaction terms and another ten possible three-way interaction terms that could be considered. If we consider even higher order interactions (e.g. four-way interactions) and quadratic terms, which can be thought of as an interaction between a variable and itself, we can see that the complexity of the model could increase dramatically. The challenge to the data analyst is not so much in fitting these models but in deciding which models to investigate in the first place. This is a topic that is returned to in the sessions on strategies of analysis later in the course.