

Module 1: Regression models

1.1	Introduction	1
1.2	A few examples	1
1.3	General regression models	5
1.3.1	Response and explanatory variables	5
1.3.2	The general regression model	10
1.3.3	Notes of caution	12
1.4	Correlation	13
1.5	Summary	16

1.1 Introduction

In this module the general idea of regression models will be presented. **Regression models** are statistical models which describe the variation in one (or more) variable(s) when one or more other variable(s) vary. Inference based on such models is known as **regression analysis**.

A few examples of situations where regression models might apply are described in Section 1.2. The examples should give you a general idea of regression models, and a flavour of the range of situations where this sort of models apply. Furthermore, they will serve as illustrations of the theory as it is developed in the course. Some terminology and a definition of the general regression model is given in Section 1.3. The concept of correlation is closely related to regression. A brief discussion of correlation is given in Section 1.4.

1.2 A few examples

Example 1.1 *Driving a car at constant speed*

Imagine that you are driving a car at 50 km/h (kilometres per hour), and consider the two variables time and distance. If you are driving at constant speed, then the theoretical relationship between time and distance covered is given by the straight line in Figure 1.1(a).

In a perfect world, where speed and distance could be measured without error, all observations would lie exactly on this straight line. However, in reality it is impossible to keep the speed exactly constant and to measure the precise distance. Therefore, in a scatterplot

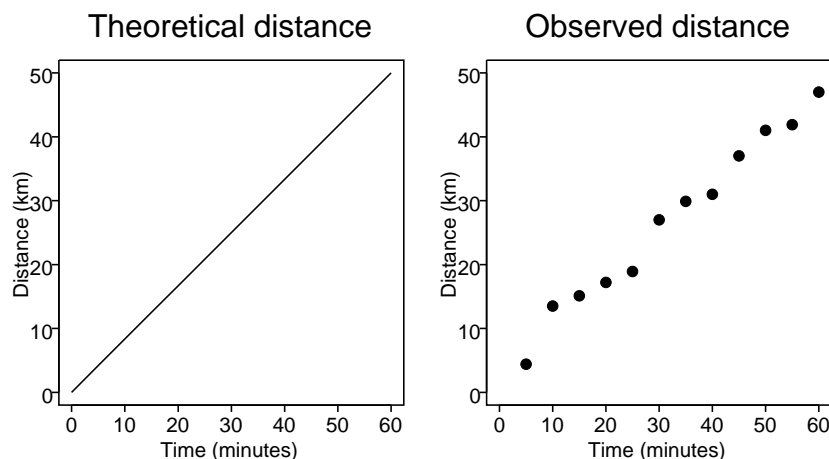


Figure 1.1: Driving at constant speed. (a) theoretical relationship. (b) scatterplot of hypothetical observations.

of ‘real’ data, the points would deviate from the theoretical straight line. A more realistic scatterplot might look something like Figure 1.1(b).

For this example, we need a model that will describe the linear relationship between the two variables, and, at the same time, take the variation away from the line into account. \diamond

Example 1.2 *Age and height of children*

Generally, a child’s height will increase with its age—up to a certain age, at least. But the growth pattern may be very different from one child to the next. Suppose you are interested in the general, overall, growth pattern of young children; one idea would be to follow a number of children over time and measure their heights at different ages. Such data would provide an indication of the overall growth pattern.

For example, as a part of a study of nutrition in developing countries, the heights of 161 children from the Egyptian village Kalama were measured each month for several years. The purpose was to find a model for the general growth pattern for later comparisons with children from other communities. The scatterplot of the data is given in Figure 1.2. Each point represents the average height of the children at the appropriate age. Not surprisingly, the scatterplot shows that height and age are closely related. Also, there seems to be a linear growth pattern, within the given age range.

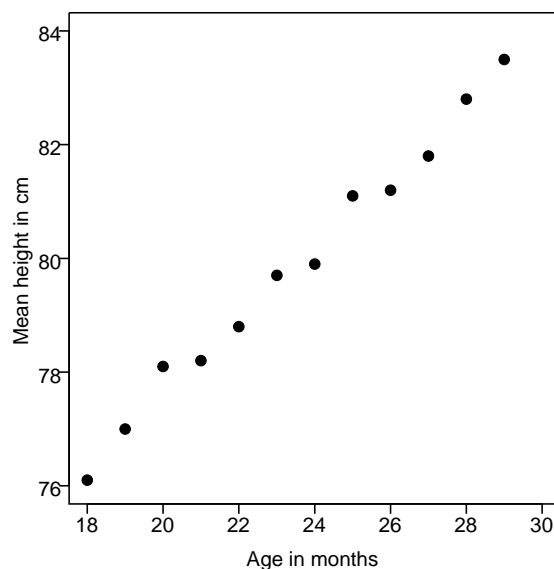


Figure 1.2: Average height and age of Egyptian children.

A suitable model for this relationship might be a straight line describing the overall growth pattern, and an error term allowing for random variation away from the line.

Further details on this dataset can be found here.



Example 1.3 *Wind power*

An investigation was made into how the direct current output from a wind power generator changes with wind speed. A scatterplot of the recorded data is shown in Figure 1.3.

The scatterplot suggests a slightly curved relationship between the current output and the wind speed - perhaps a logarithmic curve, or a square root curve. Note that there is some variation about the curve.

Further details on this dataset can be found here.



Example 1.4 *Olympic gold medal performances in high jump*

The Olympic gold medal performances in high jump for the period 1900 to 1984 are given in the scatterplot in Figure 1.4. (Note there are some missing observations due to the two World Wars.)

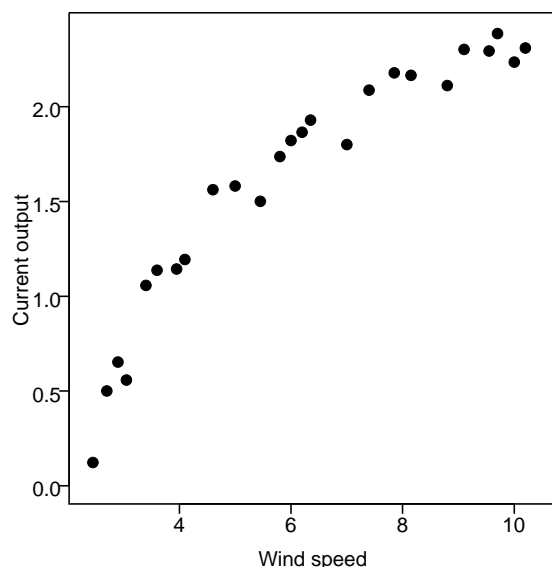


Figure 1.3: Current output against wind speed.

It appears that the Olympic gold medal performance increased almost linearly over time. Thus, a model might be a straight line describing the relationship between the performance and the year, and an error term allowing for some variation about the line.

Further details concerning this dataset can be found here.

◇

Example 1.5 *Cracking paint*

A research study was conducted on cracking of latex paint on wooden structures. The primary concern in the study was to investigate the effects of water permeability and fracture energy (energy to propagate a crack through paint film) on a given paint crack rating. Figure 1.5 shows two scatterplots of the crack rating against (a) the permeability, and (b) the fracture energy.

It seems that there is a relationship between the crack rating and the water permeability, and a relationship between the crack rating and the fracture energy. There is a lot of scatter in both scatterplots; either relationship could be straight-line or perhaps a more complicated function. A model for these data might be that the crack rating is some function of the water permeability *plus* a some function of the fracture energy *plus* an error term allowing for random variation.

Further details on this dataset can be found here.

◇

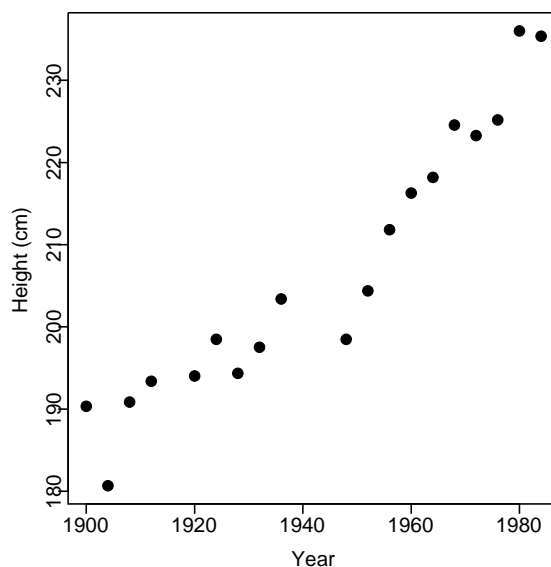


Figure 1.4: Olympic gold medal performances in high jump.

1.3 General regression models

In Subsection 1.3.1, some standard terminology is introduced before the general regression model is defined in Subsection 1.3.2. The section is concluded with Subsection 1.3.3, which briefly discusses some of the considerations one must take into account when using and interpreting regression models.

1.3.1 Response and explanatory variables

We use regression models for studying how changes in one or more variables will change the value of another variable. Generalising slightly, we can talk about a variable *‘explaining’* some of the variation in another variable. For example, in Example 1.2, we wished to ‘explain’ the height of an Egyptian child by its age, in Example 1.3, the interest was to ‘explain’ the current output of a wind mill by the wind speed, in Example 1.5, we wished to investigate how the crack rating of paint could be ‘explained’ by the water permeability and the fracture energy of the paint. In all three examples, the data were collected from studies, or experiments, specifically designed to examine how changes in one (or two, in Example 1.5) variables would affect another variable. Variables which are used to *explain* other variables are called **explanatory variables**. Thus, in Example 1.2, age is an explanatory variable, in Example 1.3, wind speed is an explanatory variable, and in Example 1.5, there are two explanatory variables: water permeability and fracture energy. In Example 1.4, data are not

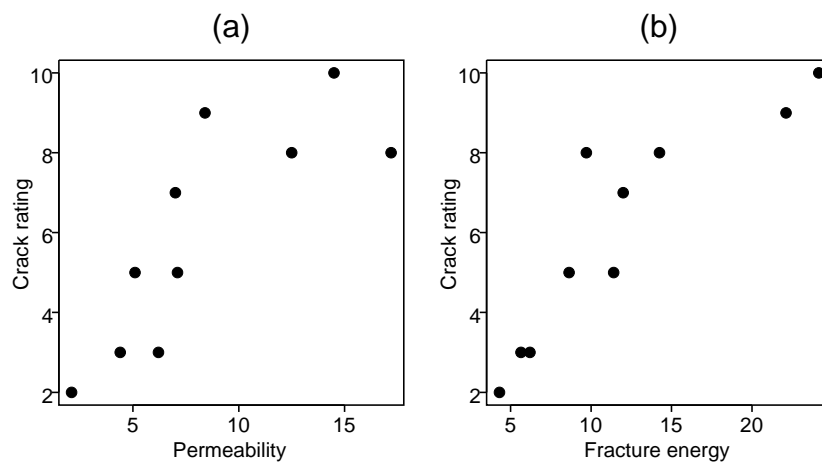


Figure 1.5: Paint crack rating against (a) water permeability and (b) fracture energy.

collected according to a pre-specified design. However, it is still the case that one variable (Olympic gold medal performance) has been recorded for different values of the other (year), and there is an interest in investigating how the performances vary as a function of the year. Thus, year can be regarded as an explanatory variable.

In each example, we wished to use the explanatory variable(s) to explain the remaining variable. The variable which is ‘explained’ is called the **response variable**. You can think of it, as *responding* to the value of the explanatory variable. For example, the wind power generator in Example 1.3 *responds* to a given wind speed with a certain current output; or, in Example 1.2 the height of an Egyptian child is a *response* to the age of the child. Thus, in Example 1.3, the current output is the response variable, and in Example 1.2, the height is the response variable. Likewise, the response variable in Example 1.4 is the Olympic gold medal performance, and the response variable in Example 1.5 is the crack rating.

Some statistical literature use different terminology for response and explanatory variables. A response variable is often called a **dependent variable**, and an explanatory variable is sometimes called an **independent variable**, or a **predictor**, or **regressor**. In this course we shall consistently use the terms response variable and explanatory variable.

Note that, occasionally, it is not clear from the data which variable is the response variable and which is the explanatory variable. It depends on the nature of the investigation. For example, the aim of the study in Example 1.2 was to find out how the height of Egyptian

children (response variable) depended on their age (explanatory variable). However, similar data could be used in an investigation on how to tell the approximate age of a child (response variable) by its height (explanatory variable). In most cases, though, it is clear from the objective of the study which is which.

For each of the next three examples, give yourself a moment to answer the following two questions before reading the comments:

- Which of the two variables is the response variable and which is the explanatory variable?
- What can you say from the scatterplot about the nature of the relationship between the variables?

Example 1.6 *Measuring mobility of elderly people*

This example concerns two methods for measuring the mobility of elderly people. The two methods are the so-called *Berg score* and *Timed Up and Go* (TUG) score. The Berg score is a measure based on how well the person performs in a number of different tasks. A low score corresponds to low mobility. The TUG score is simply the time it takes a person to get up from a chair, walk three metres and return to the chair. Measuring the Berg score is much more demanding and time-consuming than measuring the TUG score. The interest is whether the quick method (the TUG score) can be used to give a good prediction of the more thorough method (the Berg score). Figure 1.6 shows a scatterplot of the two scores for 16 individuals.

Further details on this dataset can be found here.



Comment:

The aim of the study is to try and determine the Berg score on the basis of the TUG score. Thus, the Berg score is the response variable and the TUG score is the explanatory variable. The scatterplot in Figure 1.6 shows a clear downward trend. It may be a straight line, but there is a lot of scatter present so it is difficult to give a clear answer.



Example 1.7 *Paper strength*

The scatterplot in Figure 1.7 concerns the strength of Kraft paper. (Kraft paper is a thick brown type of paper used for wrapping.) The tensile strength in *p.s.i.* (pounds per square inch) of the paper was measured against the percentage of hardwood in the batch of pulp from which the paper was produced. The tensile strength is plotted against the hardwood content in Figure 1.7.

Further details on this dataset can be found here.

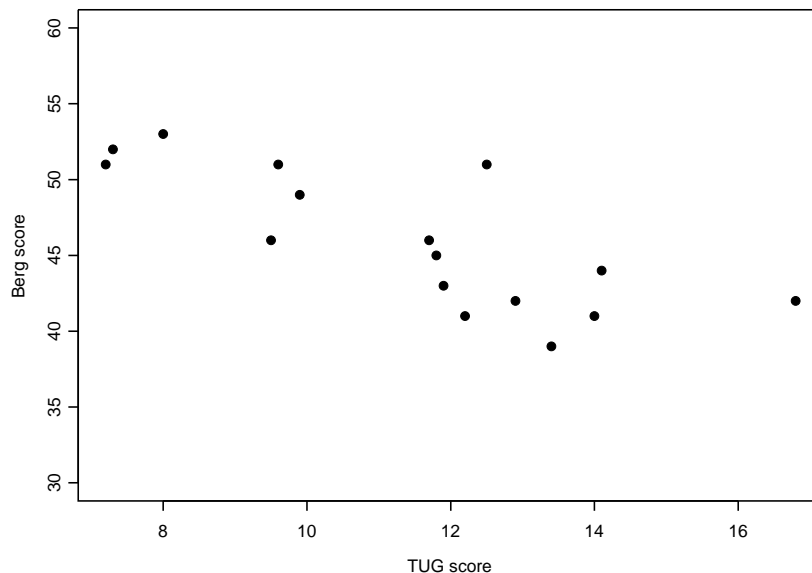


Figure 1.6: Berg-score against TUG-score.

◇

Comment:

The strength of the Kraft paper varies depending on the percentage of hardwood used in the production of the paper. Thus, the tensile strength is the response variable and the percentage of hardwood is the explanatory variable. The scatterplot in Figure 1.7 suggests a functional relationship between the tensile strength and the percentage of hardwood. Possibly a quadratic, or cubic curve, plus some random variation. It seems that a content of about 10% produces the strongest paper.

◇

Example 1.8 *When do babies crawl?*

The University of Denver Center for Infant Development designed and conducted a study in the period 1988-1991, investigating why there is much variation in the age when babies first crawl. One hypothesis was that there might be an association between the babies' first crawling age and the average ambient temperature six months after their births. (Babies usually start to crawl when they are about six months old.) The argument was that babies might take longer to learn to crawl in cold months when thick clothes restrict their movements, than in warm months. Data were collected on 208 boys and 206 girls. The parents reported birth month, and age at which their child could crawl four feet in one minute. A scatterplot of the observations are given in Figure 1.8.

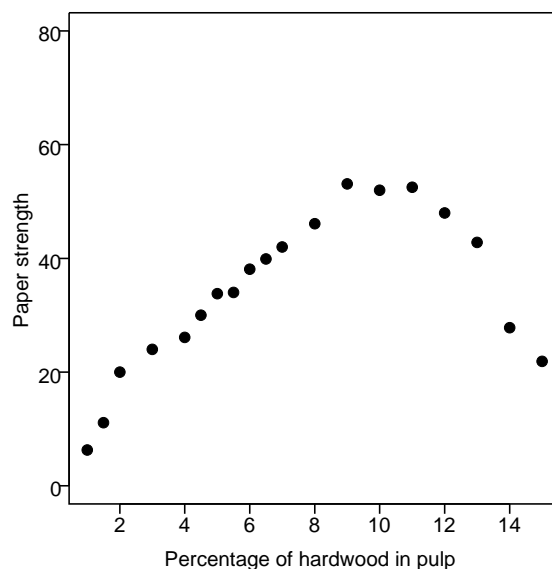


Figure 1.7: Paper strength against hardwood contents.

Further details on this dataset can be found here.

◇

Comment:

In this example, the age of crawling is the response variable, and the ambient temperature is the explanatory variable. There is a clear downward trend in the scatterplot. That is, it seems that the average crawling age is reduced when the average ambient temperature 6 months after birth increases. Thus, these data support the hypothesis in the study. There is a lot of scatter present in Figure 1.8, but there is some indication of a straight-line relationship between the two variables. However, one data point, (28.58, 11.11), lies a bit away from the general pattern. It may be an outlier.

◇

In regression, response variables are always regarded as random variables, whereas explanatory variables are usually regarded as non-random. As a consequence, all the scatter away from the main trend in a scatterplot is ascribed to the response variable, only. This assumption makes good sense in the cases where data were collected from a study specifically designed to examine how the response variable depends on an explanatory variable. For instance, in Example 1.7, several batches of Kraft paper were produced, each with a predetermined (non-random) proportion of hardwood in the pulp. The strength of the paper was

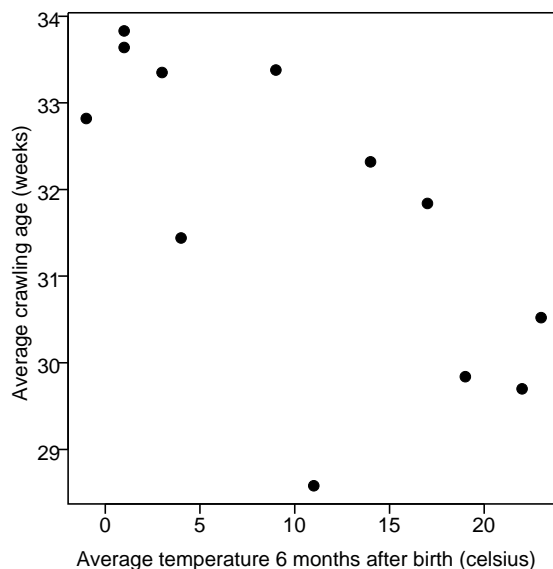


Figure 1.8: Average crawling age and the average temperature six months after birth.

then measured, introducing some random variation to the model. In Examples 1.2 and 1.4, the response variables were measured at fixed (non-random) time intervals: every month, and every year, respectively. In examples of this sort, it is clear that the explanatory variables are non-random. However, in other examples, it can seem a bit strange: in Example 1.6, the Berg score is random, while the TUG score is not! Nevertheless, we shall regard the explanatory variables as non-random throughout the course.

Since a response variable is random, we shall denote it by an upper-case letter, *e.g.* Y_i , while a non-random explanatory variable will be by a lower-case letter, *e.g.* x_i . If there is more than one explanatory variable, $x_{i,1}, \dots, x_{i,p}$, say, we let x_i denote the vector of explanatory variables $x_i = (x_{i,1}, \dots, x_{i,p})$. Note that, it is standard to always plot the response variable along the y -axis, and an explanatory variable along the x -axis in a scatterplot.

1.3.2 The general regression model

In each of Examples 1.1–1.8, we have described the relationship between the response variable and the explanatory variables on the basis of two main features of the scatterplot. In each case, there has been some sort of functional relationship (either straight-line or a curved) between the variables, and we have made some notion about the amount of scatter in the plot. The general regression model takes both of these features into account. It is defined as follows.

Suppose that we have a response variable Y and a p -dimensional vector of explanatory variables $x_i = (x_{i,1}, \dots, x_{i,p})$, then the **general regression model** is given by

$$Y_i = h(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $h(\cdot)$ represents some function, and the ε_i s are independent random variables with zero mean and constant variance for all $i = 1, \dots, n$.

The function $h(\cdot)$ is called the **regression curve**. It is the function which describes the overall trend in the scatterplot; that is, it is the function that relates the response variable to the explanatory variables. Note that $h(x)$ is non-random as it is a function of the non-random explanatory variables. The random terms ε_i , $i = 1, \dots, n$, are called **random errors**. Since the random variation in the response variable Y_i is modelled by random errors ε_i , $i = 1 \dots n$, the Y_i s are independent random variables with means $h(x_i)$, $i = 1, \dots, n$, and constant variance (equal to the variance of ε_i).

Example 1.2 (continued) *Age and height of children*

In Example 1.2, we argued that the average height (Y_i) of the Egyptian children seems to depend linearly of their ages (x_i). Thus, the following regression model might describe these data well.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 161,$$

where β_0 and β_1 are unknown parameters, and ε_i are independent random variables with zero mean and constant variance.

◇

Example 1.5 (continued) *Cracking paint*

The crack rating (Y_i) of latex paint appeared to be related to the water permeability ($x_{i,1}$), and to the fracture energy ($x_{i,2}$). A suitable regression model for these data might be of the form:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, \quad i = 1, \dots, 10,$$

where β_0 , β_1 and β_2 are unknown parameters, and ε_i are independent random variables with zero mean and constant variance.

◇

Example 1.7 (continued) *Paper strength*

The relationship between the tensile strength (Y) of Kraft paper, and the percentage of hardwood (x) in the pulp from which the paper was made, appeared to be a curve rather than a straight line. A possible model for these data might be the following.

$$Y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon_i, \quad i = 1, \dots, 19,$$

where β_0, \dots, β_3 are unknown parameters, and ε_i are independent random variables with zero mean and constant variance.

◇

In the case where there is only one explanatory variable, such as Examples 1.1–1.4, and 1.6–1.8, the regression model is called a **simple regression model**. Module 2 concerns simple regression models. In the general case, where there are two or more explanatory variables, such as Example 1.5, the general regression model is usually referred to as a **multiple regression model**. Multiple regression models are discussed in Modules 3 to 5.

In the definition of the general regression model, the function $h(\cdot)$ may be any function. However, some functions are encountered in regression models more often than others. The most common function is a straight line. That is, when $h(\cdot)$ is of the form

$$h(x_i) = h(x_{i,1}, x_{i,2}, \dots, x_{i,p}) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p},$$

where β_0, \dots, β_p are unknown parameters. Such models are called **linear regression models**. You have already seen several examples where a linear regression model might be suitable: Examples 1.1, 1.2, 1.4, 1.6 and 1.8. Most of this course is concerned with linear regression models. A special case of the linear regression model is the case when there is only one explanatory variable. That is, the function $h(\cdot)$ is on the form

$$h(x_i) = \beta_0 + \beta_1 x_i,$$

for some unknown parameters β_0 and β_1 . This model is called a **simple linear regression model**. Module 2 concern simple linear regression models.

Notice that nowhere have we assumed anything about the distribution of the response variable. We have only assumed that the response variables are independent with mean $h(x_i)$ and constant variance. However, in order to use the model for statistical inference, such as testing hypotheses about the model, or using it to make predictions on the response variable for new values of the explanatory variables, it is useful to know the distribution of the response variable. It is often appropriate to assume that Y has a normal distribution. In this course we shall almost exclusively consider models which satisfy this assumption. Sometimes, other distributions are more appropriate, for example, the gamma distribution, or the Poisson distribution. Such models, amongst others, are studied in ST112.

1.3.3 Notes of caution

In this subsection, we make two important points that one should be aware of when using regression models.

The first point concerns the validity of the model. Recall that we can often use a scatterplot of the data to give a first indication of what the function $h(\cdot)$ in a regression model might be. (As we did in Examples 1.1–1.8.) Further on in the course, you will learn a more stringent method to check whether or not a given function describes the pattern in the data well. In either case, $h(\cdot)$ is determined on the basis of the *data we have observed*. That is, any functional relationship between the response variable and the explanatory variable displayed in the scatterplot relates to our particular set of data. For data outside the range of the observations, the relationship may—in some cases—be very different.

Example 1.2 (continued) *Height and age of children*

The scatterplot in Example 1.2 indicated a straight-line relationship between the height and age of children. This relationship may well be true within the observed age-range, from 18 to 30 months, but extrapolating till the age of 20 years, say, is probably not a sound idea. The relationship will no longer be linear. For this wider age-range, we would need a more complex model to describe the relationship between the variables.

◇

One should always be careful when extrapolating a relationship. In general, one should not use a model outside the observed range.

The second point relates to the interpretation of conclusions drawn from a regression model. It is sometimes tempting to interpret a relationship in a regression model as if the explanatory variable *is the reason* for changes in the response variable. That is, that changes in the explanatory variables *cause* changes in the response variable. But a regression model does not say anything about *causation*, it simply states that if the value of the explanatory is changed, the value of the response variable also changes. There can be many reasons for an observed relationship. Of course, one possibility is that changes in the explanatory variable do cause changes in the response variable. But it is just as possible that it is the changes in the response variable which cause changes in the explanatory variable; or, that it is changes in a third variable which cause changes in both the response and the explanatory variable. Finally, it is also a possibility that the observed relationship is a coincidence; that is, the pattern in the scatterplot is simply due to random variation!

Example 1.6 (continued) *Mobility of elderly people*

In Example 1.6, we found that a person who has a high TUG score will usually have a low Berg score. But that does not mean that a high TUG score *causes* a low Berg score. Presumably, the relationship is due to the fact that a person with poor mobility will score high on the TUG score, and low on the Berg score. Hence, it is a third variable ‘mobility’ that causes changes in both the TUG and the Berg scores.

◇

In order to assert a *causal relationship*, the study must be carefully designed to rule out all other plausible explanations for the observed relationship.

1.4 Correlation

A statistical concept closely related to linear regression is that of *correlation*. **Correlation** refers to the situation where we have two random variables X and Y , and wish to measure the strength of the linear association between the two: the association is strong if knowing the value of one variable can give us a (reasonably) precise idea of the value of the other variable; the association is weak if we can only get a very rough estimate.

Note that there is an important difference between this situation and the linear regression situation: here, both X and Y are random, whereas in regression we always regard the explanatory variable x as a *non-random* variable (*c.f.* Subsection 1.3.1.). Furthermore, in linear regression we look for a *straight-line relationship* between Y and x ; here we are interested in the *strength* of the linear association between X and Y .

Example 1.2 (continued) *Age and height of children*

Suppose that we regard both age and height in Example 1.2 as random variables, and that we wish to assess the strength of the linear association between these two variables. You can see in Figure 1.9 (reproduced from Figure 1.2) that there seems to be a strong linear association between age and height: if you know the age of a child, you can get a fairly good estimate of the child's height, and *vice versa*. Further, the association is **positive** in the sense that if we increase (decrease) the value of one variable, the value of the other variable will also increase (decrease).

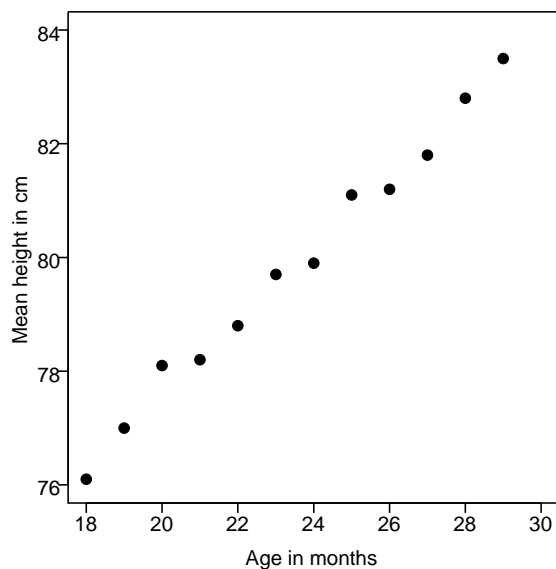


Figure 1.9: Average height and age of Egyptian children.

◇

Example 1.6 (continued) *Measuring mobility of elderly people*

We can consider the strength of the linear association between the TUG and Berg scores in Example 1.6 by regarding both variables as random. Figure 1.10 (reproduced from Figure

1.6) suggests that there is a straight-line relationship between the two variables, but the association is weaker than the linear association in Example 1.2. There is much more scatter in the plot, making the linear relationship more diffuse: by knowing the TUG score, you can only get a rough estimate of the Berg score, and *vice versa*. Note that the association is **negative** in the sense that if we increase (decrease) the value of one variable, the value of the other variable will decrease (increase).

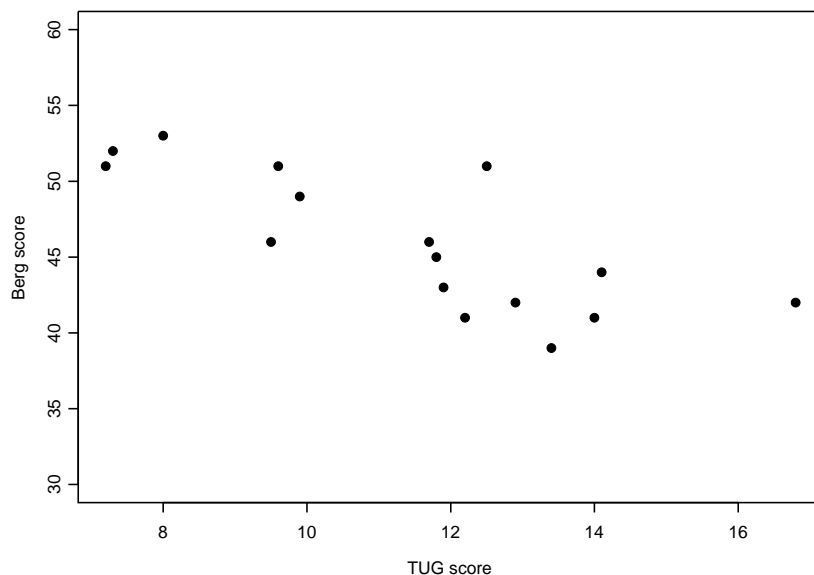


Figure 1.10: Berg-score against TUG-score.

◇

The most common measure of strength of a linear association between two variables X and Y is the **Pearson correlation coefficient** (or **correlation coefficient**, or simply **correlation**), usually denoted by ρ (rho). It is given by

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}},$$

where $\text{var}(X)$ and $\text{var}(Y)$ are the population variances of X and Y , respectively, and

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

is the **covariance** between X and Y . In particular, $\text{cov}(X, Y)$ describes the average amount by which X and Y vary with each other, or *co-vary*.

The correlation coefficient ρ can take any value between -1 and 1, inclusive. If ρ is positive, the association is positive, and if ρ is negative, the association is negative. Further, the larger the magnitude $|\rho|$ of ρ is, the stronger is the association. In particular, values of -1 and 1 indicate that the relationship between X and Y is perfectly linear (either positive, if $\rho = 1$, or negative, if $\rho = -1$). That is, all values of (X, Y) lie on the same line $y = \alpha x$, for some α . Conversely, $\rho = 0$ indicates that there is no linear relationship.

An estimate of ρ is given by the sample correlation r , that is,

$$r = \frac{s_{xy}/(n-1)}{s_{(x)}s_{(y)}},$$

where $s_{(x)} = \sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2 / (n-1)}$ and $s_{(y)} = \sqrt{\sum_{i=1}^n (y_i - \bar{y}_i)^2 / (n-1)}$ are the standard deviations for x and y , respectively, and $s_{xy}/(n-1) = \sum_{i=1}^n (x_i - \bar{x}_i)(y_i - \bar{y}_i) / (n-1)$ is the sample covariance.

Example 1.2 (continued) *Age and height of children*

The standard deviation of the age-variable (X) is $s_x = 3.6056$, the standard deviation of the height-variable (Y) is $s_y = 2.3024$, and the covariance is $\text{cov}(X, Y) = 8.2546$. Thus, the sample correlation is

$$r = \frac{8.2546}{3.6056 \times 2.3024} = 0.9943.$$

In agreement with the scatterplot of the data, this value suggests a very strong linear association between the two variables.

◇

Example 1.6 (continued) *Measuring mobility of elderly people*

The sample correlation for these data is given by

$$r = -0.7821.$$

The magnitude of the sample correlation is smaller than that in Example 1.2, indicating that the linear association is weaker for these data than for the data on heights and ages of children. Notice also that the correlation is negative, indicating that the TUG and Berg scores are negatively associated.

◇

1.5 Summary

In this module, the general regression model has been introduced. A general regression model consists of a function describing how one variable (the response variable) is related to one or more other variables (explanatory variables), and a term which models the random variation in the response variable. Several examples have been given of situations for which regression models might apply. Further, two important points have been made on the use and

interpretation of regression models: one should be careful when extrapolating relationships, and one should be careful not to over-interpret the model. Finally, the concept of correlation has been introduced. Correlation is related to linear regression models in the sense that both concepts concern straight-line relationships between variables. However, in linear regression situations the interest is to look for straight-line relationships between a random response variable and non-random explanatory variables; whereas in correlation situations the interest is on the strength of the linear association between two random variables.

Keywords: regression model, regression analysis, explanatory variable, response variable, dependent variable, independent variable, predictor, regressor, general regression model, random error, regression curve, simple regression model, multiple regression model, linear regression model, simple linear regression model, correlation, positive association, negative association, Pearson correlation coefficient, covariance.