Econometrics

Thomas Andren

bookboon
The eBook company

Thomas Andren

Econometrics

Econometrics

ISBN 978-87-7681-235-5

Contents

1. Basics of probability and statistics

The purpose of this and the following chapter is to briefly go through the most basic concepts in probability theory and statistics that are important for you to understand. If these concepts are new to you, you should make sure that you have an intuitive feeling of their meaning before you move on to the following chapters in this book.

1.1 Random variables and probability distributions

The first important concept of statistics is that of a **random experiment**. It is referred to as any process of measurement that has more than one outcome and for which there is uncertainty about the result of the experiment. That is, the outcome of the experiment can not be predicted with certainty. Picking a card from a deck of cards, tossing a coin, or throwing a die, are all examples of basic experiments.

The set of all possible outcomes of on experiment is called the **sample space** of the experiment. In case of tossing a coin, the sample space would consist of a head and a tail. If the experiment was to pick a card from a deck of cards, the sample space would be all the different cards in a particular deck. Each outcome of the sample space is called a **sample point**.

An **event** is a collection of outcomes that resulted from a repeated experiment under the same condition. Two events would be **mutually exclusive** if the occurrence of one event precludes the occurrence of the other event at the same time. Alternatively, two events that have no outcomes in common are mutually exclusive. For example, if you were to roll a pair of dice, the event of rolling a 6 and of rolling a double have the outcome (3,3) in common. These two events are therefore not mutually exclusive.

Events are said to be **collectively exhaustive** if they exhaust all possible outcomes of an experiment. For example, when rolling a die, the outcomes 1, 2, 3, 4, 5, and 6 are collectively exhaustive, because they encompass the entire range of possible outcomes. Hence, the set of all possible die rolls is both mutually exclusive and collectively exhaustive. The outcomes 1 and 3 are mutually exclusive but not collectively exhaustive, and the outcomes even and not-6 are collectively exhaustive but not mutually exclusive.

Even though the outcomes of any experiment can be described verbally, such as described above, it would be much easier if the results of all experiments could be described numerically. For that purpose we introduce the concept of a random variable. A **random variable** is a function, which assigns unique numerical values to all possible outcomes of a random experiment.

By convention, random variables are denoted by capital letters, such as X, Y, Z, etc., and the values taken by the random variables are denoted by the corresponding small letters x, y, z, etc. A random variable from an experiment can either be **discrete** or **continuous**. A random variable is discrete if it can assume only a finite number of numerical values. That is, the result in a test with 10 questions can be 0, 1, 2, …, 10. In this case the discrete random variable would represent the test result. Other examples could be the number of household members, or the number of sold copy machines a given day. Whenever we talk about random variables expressed in units we have a discrete random variable. However, when the number of unites can be very large, the distinction between a discrete and a continuous variable become vague, and it can be unclear whether it is discrete or continuous.

A random variable is said to be continuous when it can assume any value in an interval. In theory that would imply an infinite number of values. But in practice that does not work out. Time is a variable that can be measured in very small units and go on for a very long time and is therefore a continuous variable. Variables related to time, such as age is therefore also considered to be a continuous variable. Economic variables such as GDP, money supply or government spending are measured in units of the local currency, so in some sense one could see them as discrete random variables. However, the values are usually very large so counting each Euro or dollar would serve no purpose. It is therefore more convenient to assume that these measures can take any real number, which therefore makes them continuous.

Since the value of a random variable is unknown until the experiment has taken place, a probability of its occurrence can be attached to it. In order to measure a probability for a given events, the following formula may be used:

$$P(A) = \frac{\text{The number of ways event } A \text{ can occur}}{\text{The total number of possible outcomes}} \qquad (1.1)$$

This formula is valid if an experiment can result in *n* mutually exclusive and equally likely outcomes, and if *m* of these outcomes are favorable to event *A*. Hence, the corresponding probability is calculated as the ratio of the two measures: n/m as stated in the formula. This formula follows the **classical definition** of a probability.

Example 1.1
You would like to know the probability of receiving a 6 when you toss a die. The sample space for a die is {1, 2, 3, 4, 5, 6}, so the total number of possible outcome are 6. You are interested in one of them, namely 6. Hence the corresponding probability equals 1/6.

Example 1.2
You would like to know the probability of receiving 7 when rolling two dice. First we have to find the total number of unique outcomes using two dice. By forming all possible combinations of pairs we have (1,1), (1,2),…, (5,6),(6,6), which sum to 36 unique outcomes. How many of them sum to 7? We have (1,6), (2,5), (3,4), (4,3), (5,2), (6,1): which sums to 6 combinations. Hence, the corresponding probability would therefore be 6/36 = 1/6.

The classical definition requires that the sample space is finite and that the each outcome in the sample space is equally likely to appear. Those requirements are sometimes difficult to stand up to. We therefore need a more flexible definition that handles those cases. Such a definition is the so called **relative frequency definition of probability** or the empirical definition. Formally, if in *n* trials, *m* of them are favorable to the event *A*, then *P*(*A*) is the ratio *m*/*n* as *n* goes to infinity or in practice we say that it has to be sufficiently large.

Example 1.3
Let us say that we would like to know the probability to receive 7 when rolling two dice, but we do not know if our two dice are fair. That is, we do not know if the outcome for each die is equally likely. We could then perform an experiment where we toss two dice repeatedly, and calculate the relative frequency. In Table 1.1 we report the results for the sum from 2 to 7 for different number of trials.

Table 1.1 Relative frequencies for different number of trials

	Number of trials						
Sum	10	100	1000	10000	100000	1000000	∞
2	0	0.02	0.021	0.0274	0.0283	0.0278	0.02778
3	0.1	0.02	0.046	0.0475	0.0565	0.0555	0.05556
4	0.1	0.07	0.09	0.0779	0.0831	0.0838	0.08333
5	0.2	0.12	0.114	0.1154	0.1105	0.1114	0.11111
6	0.1	0.17	0.15	0.1389	0.1359	0.1381	0.13889
7	0.2	0.17	0.15	0.1411	0.1658	0.1669	0.16667

From Table 1.1 we receive a picture of how many trials we need to be able to say that that the number of trials is sufficiently large. For this particular experiment 1 million trials would be sufficient to receive a correct measure to the third decimal point. It seem like our two dices are fair since the corresponding probabilities converges to those represented by a fair die.

1.1.1 Properties of probabilities

When working with probabilities it is important to understand some of its most basic properties. Below we will shortly discuss the most basic properties.

1. $0 \leq P(A) \leq 1$ A probability can never be larger than 1 or smaller than 0 by definition.

2. If the events A, B, … are mutually exclusive we have that $P(A + B + ...) = P(A) + P(B) + ...$

Example 1.4

Assume picking a card randomly from a deck of cards. The event A represents receiving a club, and event B represents receiving a spade. These two events are mutually exclusive. Therefore the probability of the event $C = A + B$ that represent receiving a black card can be formed by $P(A+B) = P(A) + P(B)$

3. If the events A, B, … are mutually exclusive and collectively exhaustive set of events then we have that $P(A+B+...) = P(A) + P(B) + ... = 1$

Example 1.5

Assume picking a card from a deck of cards. The event A represents picking a black card and event B represents picking a red card. These two events are mutually exclusive and collectively exhaustive. Therefore $P(A+B) = P(A) + P(B) = 1$.

4. If event A and B are statistically independent then $P(AB) = P(A)P(B)$ where $P(AB)$ is called a joint probability.

5. If event A and B are not mutually exclusive then $P(A+B) = P(A) + P(B) - P(AB)$

Example 1.6

Assume that we carry out a survey asking people if they have read two newspapers (A and B) a given day. Some have read paper A only, some have read paper B only and some have read both A and B. In order to calculate the probability that a randomly chosen individual has read newspaper A and/or B we must understand that the two events are not mutually exclusive since some individuals have read both papers. Therefore $P(A+B) = P(A) + P(B) - P(AB)$. Only if it had been an impossibility to have read both papers the two events would have been mutually exclusive.

Suppose that we would like to know the probability that event A occurs given that event B has already occurred. We must then ask if event B has any influence on event A or if event A and B are independent. If there is a dependency we might be interested in how this affects the probability of event A to occur. The **conditional probability** of event A given event B is computed using the formula:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \qquad (1.2)$$

Example 1.7

We are interested in smoking habits in a population and carry out the following survey. We ask 100 people whether they are a smoker or not. The results are shown in Table 1.2.

Table 1.2 A survey on smoking

	Yes	No	Total
Male	19	41	60
Female	12	28	40
Total	31	69	100

Using the information in the survey we may now answer the following questions:

i) What is the probability of a randomly selected individual being a male who smokes?

This is just the joint probability. Using the classical definition start by asking how large the sample space is: 100. Thereafter we have to find the number of smoking males: 19. The corresponding probability is therefore: 19/100=0.19.

ii) What is the probability that a randomly selected smoker is a male?

In this case we focus on smokers. We can therefore say that we condition on smokers when we ask for the probability of being a male in that group. In order to answer the question we use the conditional probability formula (1.2). First we need the joint probability of being a smoker and a male. That turned out to be 0.19 according to the calculations above. Secondly, we have to find the probability of being a smoker. Since 31 individuals were smokers out of the 100 individuals that we asked, the probability of being a smoker must therefore be 31/100=0.31. We can now calculate the conditional probability. We have 0.19/0.31=0.6129. Hence there is 61 % chance that a randomly selected smoker is a man.

1.1.2 The probability function – the discrete case

In this section we will derive what is called the **probability mass function** or just **probability function** for a stochastic discrete random variable. Using the probability function we may form the corresponding **probability distribution**. By probability distribution for a random variable we mean the possible values taken by that variable and the probabilities of occurrence of those values. Let us take an example to illustrate the meaning of those concepts.

Example 1.8
Consider a simple experiment where we toss a coin three times. Each trial of the experiment results in an outcome. The following 8 outcomes represent the sample space for this experiment: (HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT). Observe that each sample point is equally likely to occure, so that the probability that one of them occure is 1/8.

The random variable we are interested in is the number of heads received on one trial. We denote this random variable *X*. *X* can therefore take the following values 0, 1, 2, 3, and the probabilities of occurrence differ among the alternatives. The table of probabilities for each value of the random variable is referred to as the probability distribution. Using the classical definition of probabilities we receive the following probability distribution.

Table 1.3 Probability distribution for *X*

X	0	1	2	3
P(*X*)	1/8	3/8	3/8	1/8

From Table 1.3 you can read that the probability that $X = 0$, which is denoted $P(X = 0)$, equals 1/8.

1.1.3 The cumulative probability function – the discrete case

Related to the probability mass function of a discrete random variable *X*, is its Cumulative Distribution Function, F(*X*), usually denoted CDF. It is defined in the following way:

$$F(X) = P(X \leq c) \tag{1.3}$$

Example 1.9
Consider the random variable and the probability distribution given in Example 1.8. Using that information we may form the cumulative distribution for *X*:

Table 1.4 Cumulative distribution for *X*

X	0	1	2	3
P(*X*)	1/8	4/8	7/8	1

The important thing to remember is that the outcomes in Table 1.3 are mutually exclusive. Hence, when calculating the probabilities according to the cumulative probability function, we simply sum over the probability mass functions. As an example:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

1.1.4 The probability function – the continuous case

When the random variable is continuous it is no longer interesting to measure the probability of a specific value since its corresponding probability is zero. Hence, when working with continuous random variables, we are concerned with probabilities that the random variable takes values within a certain interval. Formally we may express the probability in the following way:

$$P(a \le X \le b) = \int_a^b f(x)dx \qquad (1.4)$$

In order to find the probability, we need to integrate over the probability function, f(*X*), which is called the **probability density function**, pdf, for a continuous random variable. There exist a number of standard probability functions, but the single most common one is related to the standard normal random variable.

Example 1.10
Assume that *X* is a continuous random variable with the following probability function:

$$f(X) = \begin{cases} 3e^{-3X} & X > 0 \\ 0 & else \end{cases}$$

Find the probability $P(0 \le X \le 0.5)$. Using integral calculus we find that

$$P(0 \le X \le 0.5) = \int_0^{0.5} 3e^{-3x}dx = \left[-e^{-3x}\right]_0^{0.5} = \left[-e^{-3\times 0.5}\right] - \left[-e^{-3\times 0}\right] = -e^{-1.5} + 1 = 0.777$$

1.1.5 The cumulative probability function – the continuous case

Associated with the probability density function of a continuous random variable *X* is its **cumulative distribution function** (CDF). It is denoted in the same way as for the discrete random variable. However, for the continuous random variable we have to integrate from minus infinity up to the chosen value, that is:

$$F(c) = P(X \le c) = \int_{-\infty}^{c} f(X)dX \qquad (1.5)$$

The following properties should be noted:

1) $F(-\infty) = 0$ and $F(\infty) = 1$, which represents the left and right limit of the CDF.
2) $P(X \ge a) = 1 - F(a)$
3) $P(a \le X \le b) = F(b) - F(a)$

In order to evaluate this kind of problems we typically use standard tables, which are located in the appendix.

1.2 The multivariate probability distribution function

Until now we have been looking at univariate probability distribution functions, that is, probability functions related to one single variable. Often we may be interested in probability statements for several random variables jointly. In those cases it is necessary to introduce the concept of a **multivariate probability function**, or a **joint distribution function**.

In the discrete case we talk about the joint probability mass function expressed as

$$f(X,Y) = P(X = x, Y = y)$$

Example 1.11
Two people A and B both flip coin twice. We form the random variables X = "number of heads obtained by A", and Y = "number of heads obtained by B". We will start by deriving the corresponding probability mass function using the classical definition of a probability. The sample space for person A and B is the same and equals {(H,H), (H,T), (T,H), (T,T)} for each of them. This means that the sample space consists of 16 (4×4) sample points. Counting the different combinations, we end up with the results presented in Table 1.5.

Table 1.5 Joint probability mass function, $f(X,Y)$

		X			
		0	1	2	Total
Y	0	1/16	2/16	1/16	4/16
	1	2/16	4/16	2/16	8/16
	2	1/16	2/16	1/16	4/16
	Total	4/16	8/16	4/16	1.00

As an example, we can read that $P(X = 0, Y = 1) = 2/16 = 1/8$. Using this table we can determine the following probabilities:

$$P(X < Y) = P(X = 0, Y = 1) + P(X = 0, Y = 2) + P(X = 1, Y = 2) = \frac{2}{16} + \frac{1}{16} + \frac{2}{16} = \frac{5}{16}$$

$$P(X > Y) = P(X = 1, Y = 0) + P(X = 2, Y = 0) + P(X = 2, Y = 1) = \frac{2}{16} + \frac{1}{16} + \frac{2}{16} = \frac{5}{16}$$

$$P(X = Y) = P(X = 0, Y = 0) + P(X = 1, Y = 1) + P(X = 2, Y = 2) = \frac{1}{16} + \frac{4}{16} + \frac{1}{16} = \frac{6}{16}$$

Using the joint probability mass function we may derive the corresponding univariate probability mass function. When that is done using a joint distribution function we call it the marginal probability function. It is possible to derive a marginal probability function for each variable in the joint probability function. The marginal probability functions for X and Y is

$$f(X) = \sum_{y} f(X,Y) \text{ for all } X \tag{1.6}$$

$$f(Y) = \sum_{x} f(X,Y) \text{ for all } Y \tag{1.7}$$

Example 1.12
Find the marginal probability functions for the random variables *X*.

$$P(X=0)=f(X=0,Y=0)+f(X=0,Y=1)+f(X=0,Y=2)=\frac{1}{16}+\frac{2}{16}+\frac{1}{16}=\frac{4}{16}=\frac{1}{4}$$
$$P(X=1)=f(X=1,Y=0)+f(X=1,Y=1)+f(X=1,Y=2)=\frac{2}{16}+\frac{4}{16}+\frac{2}{16}=\frac{8}{16}=\frac{1}{2}$$
$$P(X=2)=f(X=2,Y=0)+f(X=2,Y=1)+f(X=2,Y=2)=\frac{1}{16}+\frac{2}{16}+\frac{1}{16}=\frac{4}{16}=\frac{1}{4}$$

Another concept that is very important in regression analysis is the concept of statistically independent random variables. Two random variables *X* and *Y* are said to be statistically independent if and only if their joint probability mass function equals the product of their marginal probability functions for all combinations of *X* and *Y*:

$$f(X,Y)=f(X)f(Y) \text{ for all X and Y} \qquad (1.8)$$

1.3 Characteristics of probability distributions

Even though the probability function for a random variable is informative and gives you all information you need about a random variable, it is sometime too much and too detailed. It is therefore convenient to summarize the distribution of the random variable by some basic statistics. Below we will shortly describe the most basic summary statistics for random variables and their probability distribution.

1.3.1 Measures of central tendency

There are several statistics that measure the central tendency of a distribution, but the single most important one is the expected value. The expected value of a discrete random variable is denoted E[*X*], and defined as follows:

$$E[X] = \sum_{i=1}^{n} x_i f(x_i) = \mu_X \tag{1.9}$$

It is interpreted as the mean, and refers to the mean of the population. It is simply a weighted average of all *X*-values that exist for the random variable where the corresponding probabilities work as weights.

Example 1.13
Use the marginal probability function in Example 1.12 and calculate the expected value of *X*.

$$E[X] = 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) = 0.5 + 2 \times 0.25 = 1$$

When working with the expectation operator it is important to know some of its basic properties:

1) The expected value of a constant equals the constant, $E[c] = c$
2) If c is a constant and X is a random variable then: $E[cX] = cE[X]$
3) If a, b, and c are constants and X, and Y random variables then: $E[aX + bY + c]$ = $aE[X] + bE[Y] + c$
4) If *X* and *Y* are statistically independent then and only then: $E[X, Y] = E[X]E[Y]$

The concept of expectation can easily be extended to the multivariate case. For the bivariate case we have

$$E[XY] = \sum_X \sum_Y XYf(X, Y) \tag{1.10}$$

Example 1.14
Calculate the $E[XY]$ using the information in Table 1.5. Following the formula we receive:

$$E[X,Y] = 0\times0\times\frac{1}{16} + 0\times1\times\frac{2}{16} + 0\times2\times\frac{1}{16} + 1\times0\times\frac{2}{16} + 1\times1\times\frac{4}{16} + 1\times2\times\frac{2}{16} +$$
$$2\times0\times\frac{1}{16} + 2\times1\times\frac{2}{16} + 2\times2\times\frac{1}{16} = 1$$

1.3.2 Measures of dispersion

It is sometimes very important to know how much the random variable deviates from the expected value on average in the population. One measure that offers information about that is the variance and the corresponding standard deviation. The variance of *X* is defined as

$$Var[X] = \sigma_X^2 = E\left[(X-\mu_X)^2\right] = \sum_X (X-\mu_X)^2 f(X) \tag{1.11}$$

The positive square root of the variance is the standard deviation and represents the mean deviation from the expected value in the population. The most important properties of the variance is

1) The variance of a constant is zero. It has no variability.
2) If a and b are constants then $Var(aX+b) = Var(aX) = a^2 Var(X)$
3) Alternatively we have that Var(*X*) = E[X^2] - E[*X*]^2
4) E[X^2] = $\sum_x x^2 f(X)$

Example 1.15
Calculate the variance of X using the following probability distribution:

Table 1.6 Probability distribution for *X*

X	1	2	3	4
P(*X*)	1/10	2/10	3/10	4/10

In order to find the variance for *X* it is easiest to use the formula according to property 4 given above. We start by calculating E[X^2] and E[*X*].

E[*X*] = $1\times\frac{1}{10}+2\times\frac{2}{10}+3\times\frac{3}{10}+4\times\frac{4}{10}=3$

E[X^2] = $1^2\times\frac{1}{10}+2^2\times\frac{2}{10}+3^2\times\frac{3}{10}+4^2\times\frac{4}{10}=10$

Var[*X*] = 10 – 3^2 = 1

1.3.3 Measures of linear relationship

A very important measure for a linear relationship between two random variables is the measure of covariance. The covariance of *X* and *Y* is defined as

$$Cov[X,Y] = E[(X-E[X])(Y-E(Y))] = E[XY]-E[X]E[Y] \tag{1.12}$$

The covariance is the measure of how much two random variables vary together. When two variables tend to vary in the same direction, that is, when the two variables tend to be above or below their expected value at the same time, we say that the covariance is positive. If they tend to vary in opposite direction, that is, when one tends to be above the expected value when the other is below its expected value, we have a negative covariance. If the covariance equals zero we say that there is no linear relationship between the two random variables.

Important properties of the covariance

1) $Cov[X,X] = Var[X]$
2) $Cov[X,Y] = Cov[Y,X]$
3) $Cov[cX,Y] = cCov[X,Y]$
4) $Cov[X,Y+Z] = Cov[X,Y] + Cov[X,Z]$

The covariance measure is level dependent and has a range from minus infinity to plus infinity. That makes it very hard to compare two covariances between different pairs of variables. For that matter it is sometimes more convenient to standardize the covariance so that it become unit free and work within a much narrower range. One such standardization gives us the correlation between the two random variables.

The correlation between *X* and *Y* is defined as

$$Corr(X,Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \tag{1.13}$$

The correlation coefficient is a measure for the strength of the linear relationship and range from -1 to 1.

Example 1.16
Calculate the covariance and correlation for *X* and *Y* using the information from the joint probability mass function given in Table 1.7.

Table 1.7 The joint probability mass function for *X* and *Y*

			Y		
		1	2	3	P(*X*)
X	1	0	0.1	0	0.1
	2	0.3	0.2	0.1	0.6
	3	0	0.3	0	0.3
	P(*Y*)	0.3	0.6	0.1	1.0

We will start with the covariance. Hence we have to find E[*X*,*Y*], E[*X*] and [*Y*]. We have

$$E[X] = 1\times 0.1 + 2\times 0.6 + 3\times 0.3 = 2.2$$
$$E[Y] = 1\times 0.3 + 2\times 0.6 + 3\times 0.1 = 1.8$$
$$E[XY] = 1\times 1\times 0 + 1\times 2\times 0.1 + 1\times 3\times 0 + 2\times 1\times 0.3 + 2\times 2\times 0.2 + 2\times 3\times 0.1 +$$
$$3\times 1\times 0 + 3\times 2\times 0.3 + 3\times 3\times 0 = 4$$

This gives $Cov[X,Y] = 4 - 2.2\times 1.8 = 0.04 > 0$

We will now calculate the correlation coefficient. For that we need V[*X*], V[*Y*].

$$E\left[X^2\right] = 1^2\times 0.1 + 2^2\times 0.6 + 3^2\times 0.3 = 5.2$$
$$E\left[Y^2\right] = 1^2\times 0.3 + 2^2\times 0.6 + 3^2\times 0.1 = 3.6$$
$$V[X] = E\left[X^2\right] - E[X]^2 = 5.2 - 2.2^2 = 0.36$$
$$V[Y] = E\left[Y^2\right] - E[Y]^2 = 3.6 - 1.8^2 = 0.36$$

Using these calculations we may finally calculate the correlation using (1.13)

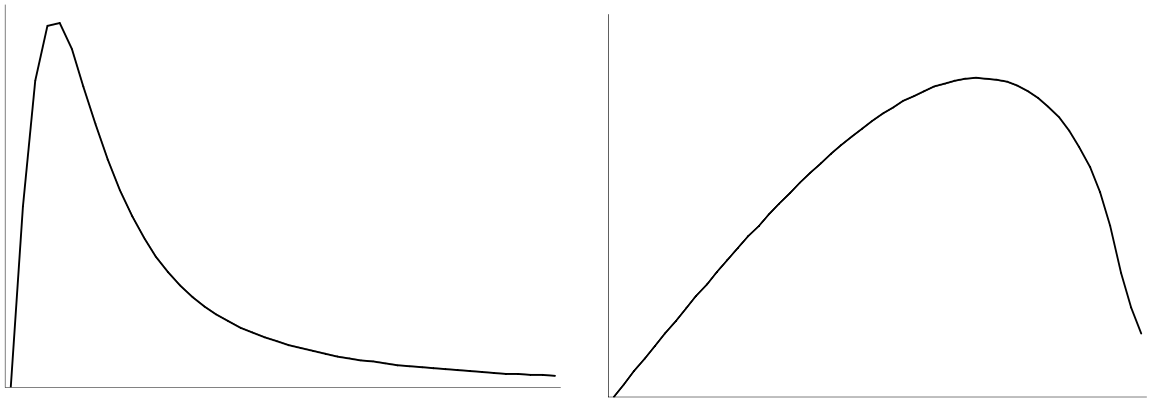
$$Corr[X,Y] = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}} = \frac{0.04}{\sqrt{0.36\times 0.36}} = 0.11$$

1.3.4 Skewness and kurtosis

The last concepts that will be discussed in this chapter are related to the shape and the form of a probability distribution function. The Skewness of a distribution function is defined in the following way:

$$S = \frac{E[X-\mu_X]^3}{\sigma_X^3} \tag{1.14}$$

A distribution can be skewed to the left or to the right. If it is not skewed we say that the distribution is symmetric. Figure 1.1 give two examples for a continuous distribution function.



a) Skewed to the right b) Skewed to the left

Figure 1.1 Skewness of a continuous distribution

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Formally it is defined in the following way:

$$K = \frac{E\left[X - \mu_X\right]^4}{\left[E\left[X - \mu_X\right]^2\right]^2} \tag{1.15}$$

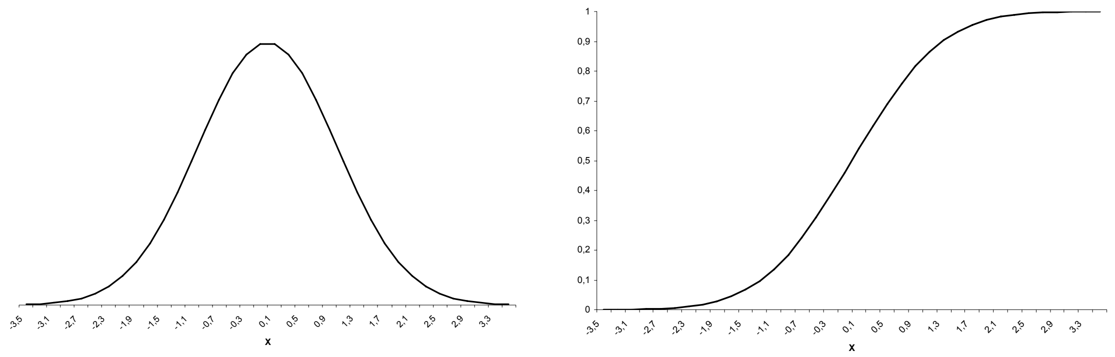
When a symmetric distribution follows the standard normal it has a kurtosis equal to 3. A distribution that are long tailed compared with the standard normal distribution has a kurtosis greater than 3 and if it is short tailed compared to the standard normal distribution it has a kurtosis that is less than three. It should be observed that many statistical programs standardize the kurtosis and presents the kurtosis as *K*-3 which means that a standard normal distribution receives a kurtosis of 0.

2. Basic probability distributions in econometrics

In the previous chapter we study the basics of probability distributions and how to use them when calculating probabilities. There exist a number of different probability distributions for discrete and continuous random variables, but some are more commonly used than others. In regression analysis and analysis related to regression analysis we primarily work with continuous probability distributions. For that matter we need to know something about the most basic probability functions related to continuous random variables. In this chapter we are going to work with the normal distribution, student t-distribution, the Chi-square distribution and the F-distribution function. Having knowledge about their properties we will be able to construct most of the tests required to make statistical inference using regression analysis.

2.1 The normal distribution

The single most important probability function for a continuous random variable in statistics and econometrics is the so called normal distribution function. It is a symmetric and bell shaped distribution function. Its Probability Density Function (PDF) and the corresponding Cumulative Distribution Function (CDF) are pictured in Figure 2.1.



a) Normal Probability Density Function b) Normal Cumulative Distribution Function

Figure 2.1 The normal PDF and CDF

For notational convenience, we express a normally distributed random variable *X* as $X \sim N(\mu_X, \sigma_X^2)$, which says that *X* is normally distributed with the expected value given by μ_X and the variance given by σ_X^2. The mathematical expression for the normal density function is given by:

$$f(X) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X} \right)^2 \right\}$$

which should be used in order to determine the corresponding CDF:

$$P(X \le c) = \int_{-\infty}^{c} f(X)dX$$

Unfortunately this integral has no closed form solution and need to be solved numerically. For that reason most basic textbooks in statistics and econometrics has statistical tables in their appendix giving the probability values for different values of c.

Properties of the normal distribution

1. The normal distribution curve is **symmetric** around its mean, μ_X , as shown in Figure 2.1a.

2. **Approximately 68 %** of the area below the normal curve is covered by the interval of plus minus one standard deviation around its mean: $\mu_X \pm \sigma_X$.

3. **Approximately 95 %** of the area below the normal curve is covered by the interval of plus minus two standard deviations around its mean: $\mu_X \pm 2\times\sigma_X$.

4. **Approximately 99.7 %** of the area below the normal curve is covered by the interval of plus minus three standard deviations around its mean: $\mu_X \pm 3\times\sigma_X$.

5. A linear combination of two or more normal random variables is also normal.

Example 2.1
If *X* and *Y* are normally distributed variables, then $Z = aX + bY$ will also be a normally distributed random variable, where *a* and *b* are constants.

6. The **skewness** of a normal random variable is **zero**.

7. The **kurtosis** of a normal random variable equals **three**.

8. A **standard normal** random variable has a mean equal to zero and a standard deviation equal to one.

9. Any normal random variable *X* with mean μ_X and standard deviation σ_X can be transformed into a **standard normal** random variable *Z* using the formula $Z = \dfrac{X-\mu_X}{\sigma_X}$.

Example 2.2
Assume a random variable *X* with expected value equal to 4 and a standard deviation equal to 8. Using this information we may transform *X* into a standard normal random variable using the following transformation: $Z = \dfrac{X-4}{8}$. It is now easy to show that *Z* has a mean equal to 0 and a variance equal to 1. That is, we have

$$E[Z] = E\left[\frac{X-4}{8}\right] = E\left[\frac{X}{8}\right] - \frac{4}{8} = \frac{1}{8}E[X] - \frac{4}{8} = 0 ,$$

$$V[Z] = V\left[\frac{X-4}{8}\right] = V\left[\frac{X}{8}\right] = \frac{1}{64}V[X] = 1$$

Since any normally distributed random variable can be transformed into a standard normal random variable we do not need an infinite number of tables for all combinations of means and variances, but just one table that corresponds to the standard normal random variable.

Example 2.3

Assume that you have a normal random variable *X* with mean 4 and variance 9. Find the probability that *X* is less than 3.5. In order to solve this problem we first need to transform our normal random variable into a standard normal random variable, and thereafter use the table in the appendix to solve the problem. That is:

$$P(X \le 3.5) = P\left(Z \le \frac{3.5-4}{3} \right) = P(Z \le -0.167)$$

We have a negative *Z* value, and the table does only contain positive values. We therefore need to transform our problem so that it adapts to the table we have access to. In order to do that, we need to recognize that the standard normal distribution is symmetric around its zero mean and the area of the pdf equals 1. That implies that $P(Z \le -0.167) = P(Z \ge 0.167)$ and that $P(Z \ge 0.167) = 1 - P(Z \le 0.167)$. In the last expression we have something that we will be able to find in the table. Hence, the solution is:

$$P(X \le 3.5) = 1 - P(Z \le 0.167) = 1 - 0.5675 = 0.4325$$

Example 2.4
Assume the same random variable as in the previous example and calculate the following probability: $P(3.5 \le X \le 4.5)$. Whenever dealing with intervals we need to split up the probability expression in two parts using the same logic as in the previous example. Hence, the probability may be rewritten in the following way:

$$P(3.5 \le X \le 4.5) = P(X \le 4.5) - P(X \le 3.5) = P\left(Z \le \frac{4.5-4}{3}\right) - P\left(Z \le \frac{3.5-4}{3}\right)$$
$$= P(Z \le 0.167) - P(Z \le -0.167)$$

In order to find the probability for this last equality we simply use the technique from the previous example.

The sampling distribution of the sample mean

Another very important concept in statistics and econometrics is the idea of a distribution of an estimator, such as the mean or the variance. It is essential when dealing with statistical inference. This issue will discussed substantially in later chapters and then in relation to estimators of the regression parameters.

The idea is quite simple. Whenever using a sample when estimating a population parameter we receive different estimate for each sample we use. That happens because of sampling variation. Since we are using different observations in each sample it is unlikely that the sample mean will be exactly the same for each sample taken. By calculating sample means from many different samples, we will be able to form a distribution of mean values. The question is whether it is possible to say something about this distribution without having to take a large number of samples and calculate their means. The answer is that we know something about that distribution.

In statistics we have a very important theorem that goes under the name **The Central Limit Theorem**. It says:

If $X_1, X_2, \ldots, X_n$ is a sufficiently large **random sample** from any population, with mean μ_X, and variance σ_X^2, then the distribution of sample means will be approximately normally distributed with $E[\overline{X}] = \mu_X$ and variance $V[\overline{X}] = \frac{\sigma_X^2}{n}$.

A basic rule of thumb says that if the sample is larger than 30 the shape of the distribution will be sufficiently close, and if the sample size is 100 or larger it will be more or less exactly normal. This basic theorem will be very helpful when carrying out tests related to sample means.

Basics steps in hypothesis testing

Assume that we would like to know if the sample mean of a random variable has changed from one year to another. In the first year we have population information about the mean and the variance. In the following year we would like to carry out a statistical test using a sample to see if the population mean has changed, as an alternative to collect the whole population yet another time. In order to carry out the statistical test we have to go through the following steps:

1) Set up the hypothesis

In this step we have to form a null hypothesis that correspond to the situation of no change, and an alternative hypothesis, that correspond to a situation of a change. Formally we may write this in the following way:

$$H_0 : \mu_X = \mu$$
$$H_1 : \mu_X \neq \mu$$

In general we would like to express the hypothesis in such a way that we can reject the null hypothesis. If we do that we will be able to say something with a statistical certainty. If we are unable to reject the null hypothesis can just say that we do not have enough statistical material to say anything about the matter. The hypothesis given above is a so called a **two sided test**, since the alternative hypothesis is expressed with an inequality. The alternative would be to express the alternative hypothesis with larger than (>) or smaller than (<), which would resulted in a **one sided test**. In most cases, you should prefer to use a two sided test since it is more restrictive.

2) Form the test function

In this step we will use the ideas that come from the Central Limit Theorem. Since we have taken a sample and calculated a mean we know that the mean can be seen as a random variable that is normally distributed. Using this information we will be able to form the following test function:

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \sim N(0,1)$$

We transform the sample mean using the population information according to the null hypothesis. That will give us a new random variable, our test function *Z*, that is distributed according to the standard normal distribution. Observe that this is true only if our null hypothesis is true. We will discuss this issue further below.

3) Choose the level of significance for the test and conclude

At this point we have a random variable *Z*, and if the sample size is larger than 100, we know how it is distributed for certain. The fewer number of observations we have, the less we know about the distribution of Z, and the more likely it is to make a mistake when performing the test. In the following discussion we will assume that the sample size is sufficiently large so that the normal distribution is a good approximation.

Since we know the distribution of *Z*, we also know that realizations of *Z* take values between -1.96 and 1.96 in 95 % of the cases (You should confirm this using Table A1 in the appendix). That is, if we take 100 samples and calculates the sample means and the corresponding test value for each sample, on average 95 % of the test values will have values within this interval, **if our null hypothesis is correct**. This knowledge will now be used using only one sample.

If we take a sample and calculate a test value and find that the test value appear outside the interval, we say that this event is so unlikely to appear (less than 5 percent in the example above) that it cannot possible come from the distribution according to the null hypothesis (it cannot have the mean stated in the null hypothesis). We therefore say that we reject the null hypothesis in favor for the alternative hypothesis.

In this discussion we have chosen the interval [-1.96;1.96] which cover 95 % of the probability distribution. We therefore say that we have chosen a 5 % **significance level** for our test, and the end points for this interval are referred to as **critical values**. Alternatively, with a significance level of 5 % there is a 5 % chance that we will receive a value that is located outside the interval. Hence there is a 5 % chance of making a mistake. If we believe this is a large probability, we may choose a lower significance level such as 1 % or 0.1 %. It is our choice as a test maker.

Example 2.5
Assume that you have taken a random sample of 10 observations from a normally distributed population and found that the sample mean equals 6. You happen to know that the population variance equals 2. You would like to know if the mean value of the population equals 5, or if it is different from 5.

You start by formulating the relevant null hypothesis and alternative hypothesis. For this example we have:

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

You know that according to the central limit theorem the sampling distribution of sample means has a normal distribution. We may therefore form the following test function:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{6-5}{\sqrt{2/10}} = 2.236$$

We know that our test function follows the standard normal distribution (has a mean equal to zero) if the null hypothesis is true. Assume that we choose a significance level of 1 %. A significance level of 1 % means that there is a 1 % chance that we will reject the null hypothesis even though the null hypothesis is correct. The critical values according to a significance level of 1 % are [-2.576; 2.575]. Since our test value is located within this interval we cannot reject the null hypothesis. We have to conclude that the mean value of the population might be 5. We cannot say that it is significantly different from 5.

2.2 The t-distribution

The probability distribution that will be used most of the time in this book is the so called t-distribution. The t-distribution is very similar in shape to the normal distribution but works better for small samples. In large samples the t-distribution converges to the normal distribution.

Properties of the t-distribution

1. The t-distribution is symmetric around its mean.
2. The mean equals zero just as for the standard normal distribution.
3. The variance equals k/(k-2), with k being the degrees of freedom.

In the previous section we explained how we could transform a normal random variable with an arbitrary mean and an arbitrary variance into a standard normal variable. That was under condition that we knew the values of the population parameters. Often it is not possible to know the population variance, and we have to rely on the sample value. The transformation formula would then have a distribution that is different from the normal in small samples. It would instead be t-distributed.

Example 2.6
Assume that you have a sample of 60 observations and you found that the sample mean equals 5 and the sample variance equals 9. You would like to know if the population mean is different from 6. We state the following hypothesis:

$$H_0 : \mu = 6$$
$$H_1 : \mu \neq 6$$

We use the transformation formula to form the test function

$$t = \frac{\overline{X} - \mu_X}{S/\sqrt{n}} \sim t_{(n-1)}$$

Observe that the expression for the standard deviation contains an *S*. *S* represents the sample standard deviation. Since it is based on a sample it is a random variable, just as the mean. The test function therefore contains two random variables. That implies more variation, and therefore a distribution that deviates from the standard normal. It is possible to show that the distribution of this test function follows the t-distribution with *n*-1 degrees of freedom, where *n* is the sample size. Hence in our case the test value equals

$$t = \frac{\overline{X} - \mu_X}{S/\sqrt{n}} = \frac{5-6}{3/\sqrt{60}} = -2.58$$

The test value has to be compared with a critical value. If we choose a significance level of 5 % the critical values according to the t-distribution would be [-2.0; 2.0]. Since the test value is located outside the interval we can say that we reject the null hypothesis in favor for the alternative hypothesis. That we have no information about the population mean is of no problem, because we assume that the population mean takes a value according to the null hypothesis. Hence, we assume that we know the true population mean. That is part of the test procedure.

2.3 The Chi-square distribution

Until now we have talked about the population mean and performed tests related to the mean. Often it is interesting to make inference about the population variance as well. For that purpose we are going to work with another distribution, the Chi-square distribution.

Statistical theory shows that the square root of a standard normal variable is distributed according to the Chi-square distribution and it is denoted χ^2, and has one degree of freedom. It turns out that the sum of squared independent standard normal variables also is Chi-squared distributed. We have:

$$Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \chi^2_{(k)}$$

Properties of the Chi-squared distribution

1. The Chi-square distribution takes only positive values
2. It is skewed to the right in small samples, and converges to the normal distribution as the degrees of freedom goes to infinity
3. The mean value equals *k* and the variance equals 2*k*, where *k* is the degrees of freedom

In order to perform a test related to the variance of a population using the sample variance we need a test function with a known distribution that incorporates those components. In this case we may rely on statistical theory that shows that the following function would work:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

where S^2 represents the sample variance, σ^2 the population variance, and n-1 the degrees of freedom used to calculate the sample variance. How could this function be used to perform a test related to the population variance?

Example 2.7

We have a sample taken from a population where the population variance a given year was $\sigma^2 = 400$. Some years later we suspect that the population variance has increase and would like test if that is the case. We collect a sample of 25 observations and state the following hypothesis:

$$H_0 : \sigma^2 = 400$$

$$H_1 : \sigma^2 > 400$$

Using the 25 observations we found a sample variance equal to 600. Using this information we set up the test function and calculate the test value:

$$\text{Test Function} = \frac{(n-1)S^2}{\sigma^2} = \frac{(25-1)\times 600}{400} = 36$$

We decided to have a significance level of 5 % and found a critical value in Table A3 equal to 36.415. Since the test value is lower than the critical value we cannot reject the null hypothesis. Hence we cannot say that the population variance has changed.

2.4 The F-distribution

The final distribution to be discussed in this chapter is the F-distribution. In shape it is very similar to the Chi-square distribution, but is a construction of a ratio of two independent Chi-squared distributed random variables. An F-distributed random variable therefore has two sets of degrees of freedom, since each variable in this ratio has its own degrees of freedom. That is:

$$\frac{\chi_m^2}{\chi_l^2} \sim F_{m,l}$$

Properties of the F-distribution

1. The F-distribution is skewed to the right and takes only positive values
2. The F-distribution converges to the normal distribution when the degrees of freedom become large
3. The square of a t-distributed random variable with *k* degrees of freedom become F-distributed: $t_k^2 = F_{1,k}$

The F-distribution can be used to test population variances. It is especially interesting when we would like to know if the variances from two different populations differ from each other. Statistical theory says that the ratio of two sample variances forms an F-distributed random variable with $n_1 - 1$ and $n_2 - 1$ degrees of freedom:

$$\frac{S_1^2}{S_2^2} \sim F_{(n_1-1)(n_2-1)}$$

Example 2.8

Assume that we have two independent populations and we would like to know if their variances are different from each other. We therefore take two samples, one from each population, and form the following hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Using the two samples we calculate the sample variances, $S_1^2 = 8.38$ and $S_2^2 = 13.14$ with $n_1 = 26$ and $n_2 = 30$. Under the null hypothesis we know that the ratio of the two sample variances is F-distributed with 25 and 29 degrees of freedom. Hence we form the test function and calculate the test value:

$$\frac{S_1^2}{S_2^2} = \frac{8.38}{13.14} = 0.638$$

This test value has to be compared with a critical value. Assume that we choose a significance level of 5 %. Using Table A4 in the appendix, we have to find a critical value for a two sided test. Since the area outside the interval should sum up to 5 %, we must find the upper critical point that corresponds to 2.5 %. If we look for that value in the table we find 2.154. We call this upper point $F_{0.025}$. In order to find the lover point we can use the following formula:

$$F_{0.975} = \frac{1}{F_{0.025}} = \frac{1}{2.154} = 0.464$$

We have therefore received the following interval: [0.464;2.154]. The test value lies within this interval, which means that we are unable to reject the null hypothesis. It is therefore quite possible that the two population variances are the same.

3. The simple regression model

It is now time to leave the single variable analysis and move on to the main issue of the book, namely regression analysis. When looking at a single variable we could describe its behavior by using any summary statistic described in the previous chapters. Most often that would lead to a mean and a variance. The mean value would be a description of the central tendency, and the variance or the standard deviation a measure of how the average observation deviates from the mean. Furthermore, the kurtosis and skewness would say something about the distributional shape around the mean. But we can say nothing about the factors that make single observations deviate from the mean.

Regression analysis is a tool that can helps us to explain in part why observations deviate from the mean using other variables. The initial discussion will be related to models that use one single explanatory factor or variable X that explains why observations from the variable Y deviate from its mean. A regression model with only one explanatory variable is sometimes called **the simple regression model**. A simple regression model is seldom used in practice because economic variables are seldom explained by just one variable. However, all the intuition that we can receive from the simple model can be used in the multiple regression case. It is therefore important to have a good understanding of the simple model before moving on to more complicated models.

3.1 The population regression model

In regression analysis, just as in the analysis with a single variable, we make the distinction between the sample and the population. Since it is inconvenient to collect data for the whole population, we usually depends our analysis on a sample. Using this sample, we try to make inference on the population, that is, we try to find the value of the parameters that correspond to the population. It is therefore important to understand the distinction between the population regression equation and the sample regression equation.

3.1.1 The economic model

The econometric model, as appose to models in statistics in general, is connected to an economic model that motivate and explains the rational for the possible relation between the variables included in the analysis. However, the economic model is only a logical description of what the researcher believes is true. In order to confirm that the made assumptions are in accordance with the reality, it is important to specify a statistical model, based on the formulation of the economic model, and statistically test the hypothesis that the economic model propose using empirical data. However, it is the economic model that allows us to interpret the parameters of the statistical model in economic terms. It is therefore very important to remember that all econometric work has to start from an economic model.

Let us start with a very simple example. Economic theory claims that there is a relationship between food consumption and disposable income. It is believed that the monthly disposable income of the household has a positive effect on the monthly food expenditures of the household. That means that if the household disposable income increases, the food expenditure will increase as well. To make it more general we claim that this is true in general, which means that when the average disposable income increase in the population, the average food expenditure will increase. Since we talk about averages we may express the economic model in terms of an expectation:

$$E[Y \mid X_1] = B_0 + B_1 X_1 \tag{3.1}$$

The conditional expectation given by (3.1) is a so called regression function and we call it the **population regression line**. We have imposed the assumption that the relationship between Y and X_1 is linear. That assumption is made for simplicity only, and later on when we allow for more variables, we may test if this is a reasonable assumption, or if we need to adjust for it. The parameters of interest are B_0 and B_1. In this text we will use capital letters for population parameters, and small letters will denote sample estimates of the population parameters. B_0 will represent the average food expenditure by households when the disposable income is zero $(X_1 = 0)$ and is usually referred to as the **intercept** or just the constant. The regression function also shows that if B_1 is different from zero and positive, the conditional mean of Y on X_1 will change and increase with the value of X_1. Furthermore, the **slope coefficient** will represent *the marginal propensity to spend on food*:

$$B_1 = \frac{dE[Y \mid X_1]}{dX_1}$$

3.1.2 The econometric model

We now have an economic model and we know how to interpret its parameters. It is therefore time to formulate the econometric model so that we will be able to estimate the size of the population parameters and test the implied hypothesis. The economic model is linear so we will be able to use linear regression analysis.

The function expressed by (3.1) represents an average individual. Hence when we collect data, individuals will typically not fall on the regression line. We might have households with the same disposable income, but with different level of food expenditures. It might even be the case that not a single observation is located on the regression line. This is something that we have to deal with. For the observer it might appear that the single observations locate randomly around the regression line. In statistical analysis we therefore control for the individual deviation from the regression line by adding a stochastic term (U) to (3.1), still under the assumption that the average observation will fall on the line. The econometric model is therefore:

$$Y_i = B_0 + B_1 X_{1i} + U_i \tag{3.2}$$

The formulation of the econometric model will now be true for all households, but the estimated population parameters will refer to the average household that is considered in the economic model. That is explicitly denoted by the subscript *i*, that appear on Y, X_1 and U but not on the parameters. We call expression (3.2) the **population regression equation**.

Adding a stochastic term may seem arbitrary, but it is in fact very important and attached with a number of assumptions that are important to fulfill. In the literature the name for the stochastic term differ from book to book and are called error term, residual term, disturbance term etc. In this text we will call the stochastic term of the population model for error term and when talking about the sample model we will refer to it as the residual term.

One important rational for the error term already mentioned is to make the equality hold true in equation (3.2) for all observations. The reason why it does not hold true in the first place could be due to omitted variables. It is quite reasonable to believe that many other variables are important determinants of the household food expenditure, such as family size, age composition of the household, education etc. There might in fact be a large number of factors that completely determines the food expenditure and some of them might be family specific. To be general we may say that:

$$Y = f(X_1, X_2, ..., X_k)$$

with *k* explanatory factors that completely determine the value of the dependent variable *Y*, where disposable income is just one of them. Hence, having access to only one explanatory variable we may write the complete model in the following way for a given household:

$$\begin{aligned} Y &= B_0 + B_1 X_1 + f(X_2, X_3, ..., X_k) \\ Y &= B_0 + B_1 X_1 + U \end{aligned}$$

Hence everything left unaccounted for will be summarized in the term *U*, which will make the equality hold true. This way of thinking of the error term is very useful. However, even if we have access to all relevant variables, there is still some randomness left since human behavior is not totally predictable or rational. It is seldom the ambition of the researcher to include everything that accounts but just the most relevant. As a rule of thumb one should try to **have a model that is as simple as possible**, and avoid including variables with a combined effect that is very small, since it will serve little purpose. The model should be a simplistic version of the reality. The ambition is never to approach the reality with the model, since that will make the model too complicated.

Sometimes it might be the case that you have received data that has been rounded off, which will make the observations for the variable less precise. Errors of measurement are therefore yet another source of randomness that the researcher sometimes has no control over. If these measurements errors are made randomly over the sample, it is of minor problem. But if the size of the error is correlated with the dependent variable you might be in trouble. In chapter 7 we will discuss this issue thoroughly.

3.1.3 The assumptions of the simple regression model

The assumptions made on the population regression equation and on the error term in particular is important for the properties of the estimated parameters. It is therefore important to have a sound understanding of what the assumptions are and why they are important. The assumptions that we will state below is given for a given observation, which means that no subscripts will be used. That is very important to remember! The assumptions must hold for each observation.

Assumption 1: $Y = B_0 + B_1 X_1 + U$

The relation between Y and X is linear and the value of Y is determined for each value of X. This assumption also impose that the model is complete in the sense that all relevant variables has been included in the model.

Assumption 2: $E[Y \mid X] = B_0 + B_1 X_1$

$E[U \mid X] = E[U] = 0$

The conditional expectation of the residual is zero. Furthermore, there must not be any relation between the residual term and the X variable, which is to say that they are uncorrelated. This means that the variables left unaccounted for in the residual should have no relationship with the variable X included in the model.

Assumption 3: $V[Y] = V[U] = \sigma^2$

The variance of the error term is homoscedastic, that is the variance is constant over different observations. Since Y and U only differ by a constant their variance must be the same.

Assumption 4: $Cov(U_i, U_j) = Cov(Y_i, Y_j) = 0 \quad i \neq j$

The covariance between any pairs of error terms are equal to zero. When we have access to a randomly drawn sample from a population this will be the case.

Assumption 5: X need to vary in the sample.

X can not be a constant within a given sample since we are interested in how variation in X affects variation in Y. Furthermore, it is a mathematical necessity that X takes at least two different values in the sample. However, we going to assume that X is fixed from sample to sample. That means that the expected value of X is the variable itself, and the variance of X must be zero when working with the regression model. But within a sample there need to be variation. This assumption is often imposed to make the mathematics easier to deal with in introductory texts, and fortunately it has no affect on the nice properties of the OLS estimators that will be discussed at the end of this chapter.

Assumption 6: U is normally distributed with a mean and variance.

This assumption is necessary in small samples. The assumption affects the distribution of the estimated parameters. In order to perform test we need to know their distribution. When the sample is larger then 100 the distribution of the estimated parameters converges to the normal distribution. For that reason this assumption is often treated as optional in different text books.

Remember that when we are dealing with a sample, the error term is not observable. That means it is impossible to calculate its mean and variance with certainty, which makes it important to impose assumptions. Furthermore, these assumptions need to hold true for each single observation, and hence using only one observation to compute a mean and a variance is impossible.

3.2 Estimation of population parameters

We have specified an economic model, and the corresponding population regression equation. It is now time to estimate the value of the population parameters. For that purpose we need a **sample regression equation**, expressed as this:

$$Y_i = b_0 + b_1 X_{1i} + e_i \tag{3.3}$$

The important difference between the population regression equation and the sample regression equation concerns the parameters and the error term. In the population regression equation the parameters are fixed constants. They do not change. In the sample regression equation the parameters are random variables with a distribution. Their mean values represent estimates of the population parameters, and their standard errors are used when performing statistical test. The error term is also an estimate and corresponds to the population error term. Sometimes it is convenient to have a separate name for the estimated error term in order to make the distinction. In this text we will call the estimated error term the **residual term**.

3.2.1 The method of ordinary least squares

There exist many methods to estimate the parameters of the population regression equation. The most common ones are the method of maximum likelihood, the method of moment and the method of Ordinary Least Squares (OLS). The last method is by all means the most popular method used in the literature and is therefore the basis for this text.

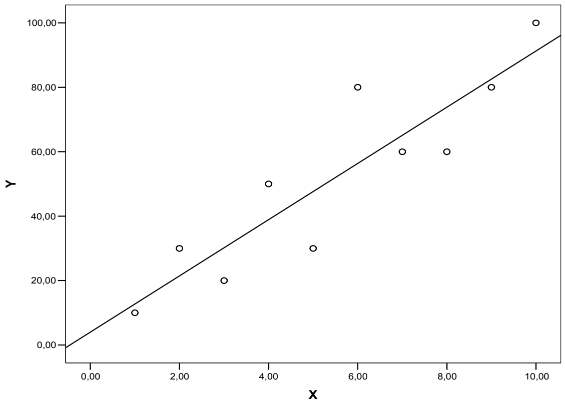


Figure 3.1 Fitted regression line using OLS

The OLS relies on the idea to select a line that represents an average relationship of the observed data similarly to the way the economic model is expressed. In Figure 3.1 we have a random sample of 10 observations. The OLS regression line is placed in such a way that the sum of the squared distances between the dots and the regression line become as small as possible. In mathematical terms using equation (3.3) we have:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right)^2 \tag{3.4}$$

In order to be more general we assume a sample size of *n* observations. The objective is to minimize the Residual Sum of Squares (RSS) expressed in (3.4) with respect to b_0 and b_1. Hence, this is a standard optimization problem with two unknown variables that is solved by taking the partial derivatives with respect to b_0 and b_1, put them equal to zero, and then solving the resulting linear equations system with respect to those two variables. We have:

$$\frac{\partial RSS}{\partial b_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right) = 0 \tag{3.5}$$

$$\frac{\partial RSS}{\partial b_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i}\right)\left(- X_{1i}\right) = 0 \tag{3.6}$$

By rearranging these two equations we obtain the equation system in normal form:

$$nb_0 + b_1\sum_{i=1}^{n} X_{1i} = \sum_{i=1}^{n} Y_i$$

$$b_0\sum_{i=1}^{n} X_i + b_1\sum_{i=1}^{n} X_{1i}^2 = \sum_{i=1}^{n} X_{1i}Y_i$$

Solving for b_0 and b_1 gives us:

$$b_0 = \overline{Y} - b_1\overline{X}_1 \tag{3.7}$$

$$b_1 = \frac{\sum_{i=1}^{n} X_{1i}Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_{1i}^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \frac{Cov\left(X_1, Y\right)}{Var(X_1)} \tag{3.8}$$

The slope coefficient b_1 is simply a standardized covariance, with respect to the variation in X_1.The interpretation of this ratio is simply: when X_1 increases by 1 unit, Y will change by b_1 units. Remember that b_0 and b_1 are random variables, and hence it is important to know how their expected values and variances look like. Below we will derive the expected value and variance for both the intercept and the variance. The variance of the intercept is slightly more involved, but since text books in general avoid showing how it could be done we will do it here, even though the slope coefficient is the estimate of primary interest.

In order to find the expected value and the variance it is convenient to rewrite the expression for the estimators in such a way that they appear to be functions of the sample values of the dependent variable Y. Since the intercept is expressed as a function of the slope coefficient we will start with the slope estimator:

$$b_1 = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)Y_i}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2} = \sum_{i=1}^{n}\underbrace{\left[\frac{\left(X_{1i} - \overline{X}_1\right)}{\sum_{i=1}^{n}\left(X_{i1} - \overline{X}_1\right)^2}\right]}_{W_i} Y_i = \sum_{i=1}^{n} W_iY_i$$

$$b_1 = \sum_{i=1}^{n} W_iY_i \tag{3.9}$$

For the intercept we do the following:

$$b_0 = \overline{Y} - b_1\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{X}_1 \underbrace{\sum_{i=1}^{n} W_i Y_i}_{(3.9)} = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)\left(B_0 + B_1 X_{1i} + U_i\right) =$$

$$= B_0 \underbrace{\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)}_{=1} + B_1 \underbrace{\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)X_{1i}}_{=0} + \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)U_i$$

Hence

$$b_0 = B_0 + \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)U_i \tag{3.10}$$

Hence, the OLS estimators are weighted averages of the dependent variable, holding in mind that W_i is to be treated as a constant. Having the OLS estimators in this form we can easily find the expected value and variance:

The expected value of the OLS estimators

$$E[b_0] = B_0 + \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)E[U_i] = B_0 \tag{3.11}$$

$$E[b_1] = E\left[\sum_{i=1}^{n} W_i Y_i\right] = E\left[\sum_{i=1}^{n} W_i\left(B_0 + B_1 X_{1i} + U_i\right)\right] = E\left[B_0 \underbrace{\sum_{i=1}^{n} W_i}_{=0}\right] + E\left[B_1 \underbrace{\sum_{i=1}^{n} W_i X_{1i}}_{=1}\right] + E\left[\sum_{i=1}^{n} W_i U_i\right]$$

and

$$E[b_1] = B_1 + \sum_{i=1}^{n} W_i \underbrace{E[U_i]}_{=0} = B_1 \tag{3.12}$$

Hence, the mean value of the sample estimators equals the population parameters. You should confirm these steps by your self. The result from the second line comes from the regression assumptions. Also remember that the population parameter is a constant and that the expected value of a constant is the constant itself. The derivation of the variance will start with the expression established at the second line above.

The variance of the OLS estimators

When deriving the variance for the intercept, we utilize the definition of the variance that is expressed in terms of expectations. We have the expected value of the squared difference, and thereafter substitute

$$V[b_0] = E[b_0 - E(b_0)]^2 = E\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}_1 W_i\right)U_i\right]^2$$

Square the expression and take the expectation and end up with

$$V\left[b_0\right]=\left(\frac{1}{n}+\frac{\overline{X}_1^2}{\sum_{i=1}^{n}\left(X_{1i}-\overline{X}_1\right)^2}\right)\sigma^2=\frac{\sigma^2\sum_{i=1}^{n}X_{1i}^2}{\sum_{i=1}^{n}\left(X_{i1}-\overline{X}_1\right)^2} \qquad (3.13)$$

Try to work out the expressions and remember that $E\left[U_i^2\right]=E\left[U^2\right]=\sigma^2$ and that $E\left[U_iU_j\right]=Cov\left[U_i,U_j\right]=0$.

$$V\left[b_1\right]=V\left[B_1+\sum_{i=1}^{n}W_iU_i\right]=V\left[\sum_{i=1}^{n}W_iU_i\right]=\sum_{i=1}^{n}W_i^2V\left[U_i\right]=\sigma^2\sum_{i=1}^{n}\left[\frac{(X_{1i}-\overline{X}_1)}{\sum_{i=1}^{n}(X_{1i}-\overline{X}_1)^2}\right]^2=\sigma^2\frac{\sum_{i=1}^{n}\left(X_{1i}-\overline{X}_1\right)^2}{\left[\sum_{i=1}^{n}(X_{1i}-\overline{X}_1)^2\right]^2}$$

and therefore

$$V\left[b_1\right]=\frac{\sigma^2}{\sum_{i=1}^{n}(X_{1i}-\overline{X}_1)^2} \qquad (3.14)$$

The covariance between the two OLS estimators can be received using the covariance operator together with expressions (3.9) and (3.10). Try it out. The covariance is given by the following expression:

$$Cov(b_0, b_1) = \frac{-\bar{X}_1\sigma^2}{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2}$$

In order to understand all the steps made above you have to make sure you remember how the variance operator works. Go back to chapter 1 and repeat if necessary. Also remember that the variance of the population error term is constant and the same over observations. If that assumption is violated we will end up with something else.

Observe that the variance of the OLS estimators is a function of the variance of the error term of the model. The larger the variance of the error term, the larger becomes the variance of the OLS estimator. This is true for the variance of the intercept, variance of the slope coefficient and for the covariance between slope and the intercept. Remember that the variance of the error term and the variance of the dependent variable coincide. Also note that the larger the variation in *X* is, the smaller become the variance of the slope coefficient. Think about that. Increased variation in *Y* has of course the opposite effect, since the variance in *Y* is the same as the variance of the error term.

The variance of the population error term, σ^2 , is usually unknown. We therefore need to replace it by an estimate, using sample information. Since the population error term is unobservable, one can use the estimated residual to find an estimate. We start by forming the residual term

$$e_i = Y_i - b_0 - b_1 X_{1i}$$

We observe that it takes two estimates to calculate its value which implies a loss of two degrees of freedom. With this information we may use the formula for the sample variance. That is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Observe that we have to divide by *n*-2, which referees to the degrees of freedom, which is the number of observations reduced with the number of estimated parameters used in order to create the residual. It turns out that this is an unbiased estimator of the population variance and it is decreasing as the number of observations increases.

3.2.2 Properties of the least squares estimator

The OLS estimator is attached to a number of good properties that is connected to the assumptions made on the regression model which is stated by a very important theorem; the **Gauss Markov theorem**.

The Gauss Markov Theorem

When the first 5 assumptions of the simple regression model are satisfied the parameter estimates are unbiased and have the smallest variance among other linear unbiased estimators. The estimators are then called **BLUE** for Best Linear Unbiased Estimators.

The OLS estimators will have the following properties when the assumptions of the regression function are fulfilled:

1) The estimators are unbiased

That the estimators are unbiased means that the expected value of the parameter equals the true population value. That means that if we take a number of samples and estimate the population parameters with these samples, the mean value of those estimates will equal the population value when the number of samples goes to infinity. Hence, on average we would be correct but it is not very likely that we will be exactly right for a given sample and a given set of parameters.

Unbiased estimators implies that

$$E[b_0] = B_0$$
$$E[b_1] = B_1$$

2) Minimum variance: Efficiency of unbiased estimators

When the variance is best, it means that it is efficient and that no other linear unbiased estimator has a better precision (smaller variance) of their estimators. It requires that the variance is homoscedastic and that it is not autocorrelated over time. Both these two issues will be discussed in chapter 9 and 10.

3) Consistency

Consistency is another important property of the OLS estimator. It means that when the sample size increase and goes to infinity, the variance of the estimator has to converge to zero and the parameter converge to the population parameters. An estimator can be biased and still consistent but it is not possible for an estimator to be unbiased and inconsistent.

4) Normally distributed parameters

Since the parameters are weighted averages of the dependent variable they can be treated as a means. According to the central limit theorem, the distribution of means is normally distributed. Hence, the OLS estimators are normally distributed in sufficiently large samples.

4. Statistical inference

Statistical inference is concerned with the issue of using a sample to say something about the corresponding population. Often we would like to know if a variable is related to another variable, and in some cases we would like to know if there is a causal relationship between factors in the population. In order to find a plausible answer to these questions we need to perform statistical test on the parameters of our statistical model. In order to carry out tests we need to have a test function and we need to know the sampling distribution of the test function.

In the previous chapter we saw that the estimators for the population parameters were nothing more than weighted averages of the observe values of the dependent variable. That is true for both the intercept and the slope coefficient. Furthermore, the distribution of the dependent variable coincides with the error term. Since the error term by assumption is normally distributed, the dependent variable will be normally distributed as well.

According to statistical theory we know that a linear combination of normally distributed variables is also normally distributed. That implies that the distribution of the two OLS estimators is normally distributed with a mean and a variance. In the previous chapter we derived the expected value and the corresponding variance for the estimators, which implies that we have all the information we need about the sampling distribution for the two estimators. That is, we know that:

$$b_0 \sim N\left(B_0, \frac{\sigma^2 \sum X_{1i}^2}{\sum \left(X_{1i} - \bar{X}_1\right)^2}\right) \tag{4.1}$$

$$b_1 \sim N\left(B_1, \frac{\sigma^2}{\sum \left(X_{1i} - \bar{X}_1\right)^2}\right) \tag{4.2}$$

Just as for a single variable, the OLS estimators works under the central limit theorem since they can be treated as means (weighted averages) calculated from a sample. When taking the square root of the estimated variances we receive the corresponding standard deviations. However, in regression analysis we call them **standard errors of the estimator** instead of standard deviations. That is to make it clear that we are dealing with a variation that is due to a sampling error. Since we use samples in our estimations, we will never receive estimates that exactly equal the corresponding population parameter. It will almost always deviate to some extent. The important point to recognize is that this error on average will be smaller the larger the sample become, and converge to zero when the sample size goes to infinity. When an estimator behaves in this way we say that the estimator is consistent as described in the previous chapter.

4.1 Hypothesis testing

The basic steps in hypothesis testing related to regression analysis are the same as when dealing with a single variable, described in earlier chapters. The testing procedure will therefore be described by an example.

Example 4.1
Assume the following population regression equation:

$$Y = B_0 + B_1 X + U \tag{4.3}$$

Using a sample of 200 observations we received the following regression results:

$$\begin{aligned} E[Y \mid X] &= 4.233 + 0.469X \\ &\quad\ (3.26)\ \ (0.213) \end{aligned} \tag{4.4}$$

The regression results in (4.4) present the regression function with the estimated parameters together with the corresponding standard errors within parentheses. It is now time to ask some questions: Has *X* any effect on *Y*? In order to answer that question we would like to know if the parameter estimate for the slope coefficient is significantly different from zero or not. We start by stating the hypothesis:

$$\begin{aligned} &H_0 : B_1 = 0 \\ &H_1 : B_1 \neq 0 \end{aligned} \tag{4.5}$$

In order to test this hypothesis we need to form the test function relevant for the case. We know that the sample estimator is normally distributed with a mean and a standard error. We may therefore transform the estimated parameter according to the null hypothesis and use that transformation as a test function. Doing that we receive:

$$\text{Test function:} \quad t = \frac{b_1 - B_1}{se(b_1)} \sim t_{n-k} \qquad (4.6)$$

The test function follows a **t-distribution** with *n-k* degrees of freedom, where *n* is the number of observations and *k* the number of estimated parameters in the regression equation (2 in this case). It takes a t-distribution since the standard error of the estimated parameter is unknown and replaced by an estimate of the standard error. This replacement increases the variation of the test function compared to what had been the case otherwise. Had the standard error been known, the test function would have been normal. However, since the number of observations is sufficiently large, the extra variation will not be of any significant importance. If the null hypothesis is true the mean of the test function will be zero. If that is not the case the test function will receive a large value in absolute terms. Let us calculate the test value using the test function:

$$\text{Test value:} \quad t = \frac{0.469 - 0}{0.213} = 2.2 \qquad (4.7)$$

The final step in the test procedure is to find the critical value that the test value will be compared with. If the test value is larger than the critical value in absolute terms we reject the null hypothesis. Otherwise, we just accept the null hypothesis and say that it is possible that the population parameter is equal to zero. In order to find the critical value we need a significance level, and it is you as a test maker that set this level. In this example we choose the significance level to be at the 5 % level. Since the degrees of freedom equals 198 the critical value found in most tables for the t-distribution will coincide with the critical value taken from the normal distribution table. In this particular case we receive:

$$\text{Critical value:} \quad t_c = 1.96 \qquad (4.8)$$

Since the test value is larger than the critical value we reject the null hypothesis and claim that there is a positive relation between *X* and *Y*.

4.2 Confidence interval

An alternative approach to the test of significance approach described by the example in the previous section is the so called confidence interval approach. The two approaches are very much related to each other. The idea is to create an interval estimate for the population parameter instead of working with a point estimate.

The basic steps are about the same. All tests need to start from some hypothesis and we will use (4.5) in this example. By choosing a 5 % significance level and the corresponding critical values from the t-distribution table we may form the following interval:

$$P(-1.96 \leq t \leq 1.96) = 95\% \qquad (4.9)$$

This expression says that there is a 95 % chance that a t-value with 198 degrees of freedom lies between the limits given by the interval. In order to form a confidence interval for our case we substitute the test function (4.6) into (4.9). That will result in the following expression:

$$P\left(-1.96 \leq \frac{b_1 - B_1}{se(b_1)} \leq 1.96\right) = 0.95$$

which may be transformed in the following way:

$$P\big(b_1 - 1.96 \times se(b_1) \leq B_1 \leq b_1 + 1.96 \times se(b_1)\big) = 0.95$$

which provides a 95 % confidence interval for B_1. Hence, there is a 95 % chance that the given interval will cover the true population parameter value. Alternatively, in repeated sampling the interval will cover the population parameter in 95 cases out of 100 on average.

If the confidence interval does not cover the value given by the null hypothesis (zero in this case) we will be able to reject the null hypothesis. By plugging in the values we receive a confidence interval that may be expressed in the following way:

$$b_1 \pm t_c \times se(b_1)$$

which in this case equals

$$0.469 \pm 1.96 \times 0.213$$

Remember that, since both the parameter and the corresponding standard errors are estimates based on sample information, the interval is random. One should therefore not forget that it is the interval that with a certain probability will cover the true population parameter, and not the other way around.

Two important concepts to remember and distinguish in these circumstances are the confidence level and significance level. They are defined in the following way:

Confidence level

The percent of the times that the constructed confidence interval will include the population parameter. When it is expressed as a percent, it is sometimes called the confidence coefficient.

Significance level

The probability of rejecting a true null hypothesis.

Hence, before being able to construct a confidence interval we have to pick a significance level, which is usually set to 5 percent. Given the significance level we know that the confidence level of our test or corresponding interval will be 95 percent. The significance level is often denoted with the Greek letter α, which implies that the confidence level equals 1-α.

4.2.1 P-value in hypothesis testing

Most econometric software that produces regression output report p-values related to each estimated parameter. To investigate the p-value is a fast way to reach the conclusion that we otherwise would receive by carrying out all the steps in the test of significance approach or the confidence interval approach. By looking at the p-value we can directly say if the parameter is significantly different from zero or not.

The P-value for sample outcome

The P-value for a sample outcome is the probability that the sample outcome could have been more extreme than the observed one.

If the P-value equals or is greater then the specified significance level: H_0 is concluded.

If the P-value is less than the specified significance level: H_1 is concluded.

Table 4.1 Regression output from Excel

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9.333333	9.0973966	1.0259345	0.334940	-11.645314	30.311981
X	7.121212	1.4661782	4.8569893	0.001260	3.7401968	10.502227

In Table 4.1 we have an example of how regression output could look like. This particular output is generated using MS Excel, but most statistical software offers this information in their output. The example is based on a random sample of 10 observations.

Observe that regression output always assumes a two sided test. That has implications on the P-value. The P-Value in this particular case can therefore be calculated as

$$P(|t| \geq 1.0259...) = P(t \leq -1.0259...) + P(t \geq 1.0259...) = 0.3349...$$

The P-value from a one-sided hypothesis would therefore be

$$P(t \geq 1.0259...) = 0.3349.../2 = 0.1675$$

Since the P-value for the intercept is larger than any conventional significance levels, say 5 percent, we can not reject the null hypothesis that the intercept is different from zero. For the slope coefficient on the other hand the P-value is much smaller than 5 percent and therefore we can reject the null hypothesis and say that it is significantly different from zero.

4.3 Type I and type II errors

Whenever working with statistical tests there is a chance that the conclusion from the test could be wrong. We could accept a false null hypothesis and we could reject a correct null hypothesis. Hence, in order for decision rules or tests of hypothesis to be good, they must be designed so as to minimize the errors of decision. For a given sample size this is a difficult task, since any attempt to minimize one of them results in increasing the other kind. The only way to reduce the chance of both types is by increasing the sample size. The two types of errors mentioned above are referred to as the type I error and the type II error.

The type I error

The probability to reject a correct null hypothesis.

$$P(\text{Reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ is true}) = \alpha$$

The type II error

The probability to accept a false null hypothesis

$$P(\text{Accept } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ is false}) = \beta$$

An additional concept related to the two types of errors is the so called power of the test. The power of the test is the probability to identify a false null hypothesis. It is always the case that we would like the power of the test to be as large as possible. However, since we need to know the true value of the population parameter we will never be able to calculate the type II error and the corresponding power of the test in practice. But for a given sample we must remember that the smaller we choose the significance level, the larger become the type II error and the smaller become the power of the test.

The power of a test

The probability to reject a false null hypothesis

$$P(\text{Reject H}_0 \mid \text{H}_0 \text{ is false}) = 1 - P(\text{Accept H}_0 \mid \text{H}_0 \text{ is false}) = 1 - \beta$$

Table 4.2 Errors in hypothesis testing

True state of nature	Researcher's decision	
	Accept H_0	Reject H_0
H_0 is true	1-α The confidence level	α (Type I error) The significance level
H_0 is false	β (Type II error)	1-β The power

Example 4.2

To better understand the mechanism of the two types of error we consider an example were we calculate the probabilities for each cell given in Table 4.2. Let us use the regression results from Example 4.1 and focus on the slope coefficient.

In that example we had a significance level of 5 percent. It was our choice, and hence we have specified that the probability to reject a correct null hypothesis should be 5 percent. Given the significance level we can calculate the interval that will be used as a decision rule for our test. If the estimated parameter is within the interval we will accept the null hypothesis, and if it is located outside the interval we will reject the null hypothesis.

In Example 4.1 the null hypothesis said that the population parameter equals zero. Together with the estimated standard error we know the distribution of the estimated parameter when the null hypothesis is correct. That is

$$b_1 \sim N(0, V(b_1))$$

Hence, an interval that covers 95 percent of the distribution is given by the endpoints of the confidence interval. For this particular case we have the following interval:

$$[-1.96 \times se(b_1), 1.96 \times se(b_1)] = [-0.41748, 0.41748] \tag{4.10}$$

Therefore, if our estimator comes from a distribution with mean zero, there is a 95 percent chance that it will be located in the above mentioned interval. The interval can therefore be seen as a decision rule. If the estimated parameter value takes a value within this interval, we should conclude that it comes from a distribution with mean zero. Since we decided about the significance level we know the probability of the type I error, since they always coincide. We will therefore go on and calculate the type II error.

In order to be able to calculate the type II error we need to know the true value, that is, we need to know the population value of the parameter. That will never happen in reality, but by assuming different values one can receive a picture of the size of the possible chance of decision error. In this example we will assume that we know the value and that it equals 0.25. Given this value we have a new distribution for our estimator that we will use when calculating the probability of a type II error, namely:

$$b_1 \sim N(0.25, V(b_1)) \tag{4.11}$$

This is the distribution related to the alternative hypothesis. In order to find the probability of a type II error we simply calculate how large part of this distribution that overlap the region of our decision rule given by interval (4.10). That is, we have to calculate the following probability using the distribution given by (4.11):

$$P(-0.41748 \le b_1 \le 0.41748) = \beta$$

In order to calculate this probability we have to use the t-value transformation and use the table for the t-distribution to find the probability. The following steps need to be done:

$$P\left(\frac{-0.41748-0.25}{0.213} \le t \le \frac{0.41748-0.25}{0.213}\right) = P(-3.134 \le t \le 0.786) = P(t \le 0.786) - P(t \le -3.134) =$$

$$0.783 - 0.0009 = 0.7821$$

Hence, with this setup there is a 78 percent chance of committing a type II error. The only way to reduce this probability is to increase the number of observations in the sample. If that is not possible you are stuck with a problem. Observe that if you decide to decrease the probability of the type I error further, the interval given by (4.10) will be even wider. When that happens, a larger portion of the true distribution will be covered, and hence increase the type II error.

Once the type II probability is calculated it is straight forward to calculate the power of the test. In this case the power of the test would be $(1-\beta)$ =1-0.7821=0.2179. Hence, there is only a 22 percent chance to reject a false null hypothesis.

4.4 The best linear predictor

Another important use of the regression model is to predict the size of the dependent variable for different values of *X*. Let us start with a definition:

Prediction and Forecasting

To make a statement about an event before the event occurs. In econometrics a statement made in advance about the value of a dependent variable using regression results.

The words prediction and forecasting are going to be used interchangeably. However, often the word prediction is used for models that covers cross sectional analysis, while predictions made using times series models on future events are called forecasting. Since the literature does not show any consensus on this part we will treat them synonymously in this text.

Assume the following population regression equation:

$$Y_t = B_0 + B_1 X_t + U_t, \quad t = 1,...,T$$

and we would like to make predictions about the future, that is we would like to know the value of *Y* in period *T*+1. We have basically two important cases to consider: known values of *X* or values of *X* with uncertainty. Whether we have exact information about *X* or not will affect the variance for the predicted value. We will start the discussion assuming that the *X* value is known, and later relax this assumption to explore the difference. The exact value of the population parameters is never an issue, and it is therefore obvious that they have to be estimated.

The predicted value of the dependent variable is therefore given by the conditional expectation of the dependent variable and is denoted in the following way:

$$E[Y_t \mid X_t] = \hat{Y}_t = b_0 + b_1 X_t$$

where the population parameters has been replaced by the sample estimators. Since the sample estimators are the same for all *t*, it is the value of X_t that generates the forecast for Y_t. Hence the forecast value of *Y* in period *T*+1 is therefore given by:

$$\hat{Y}_{T+1} = b_0 + b_1 X_{T+1}$$

This is often called a point prediction. In order to make inference on the future, we need an interval prediction as well, that is, we need to calculate the forecast error. The forecast error will help us say something about how good the prediction is. The forecast error is the difference between the predicted value and the actual value and may be expressed in the following way:

$$\begin{aligned} Y_{T+1} - \hat{Y}_{T+1} &= \left(B_0 + B_1 X_{T+1} + U_{T+1}\right) - \left(b_0 + b_1 X_{T+1}\right) \\ &= \left(B_0 - b_0\right) + \left(B_1 - b_1\right) X_{T+1} + U_{T+1} \end{aligned}$$

Now, what is the expected value of the forecast error:

$$E\left[Y_{T+1} - \hat{Y}_{T+1}\right] = \underbrace{\left(B_0 - E[b_0]\right)}_{=0} + \underbrace{E\left[\left(B_1 - b_1\right) X_{T+1}\right]}_{=0} + \underbrace{E\left[U_{T+1}\right]}_{=0} = 0$$

Since the expected value of the forecast error is zero, we have an unbiased forecast. Assuming that *X* is known, the variance of the forecast error is given by:

$$\begin{aligned} E\left[Y_{T+1} - \hat{Y}_{T+1}\right]^2 &= E\left[\left(B_0 - b_0\right) + \left(B_1 - b_1\right) X_{T+1} + U_{T+1}\right]^2 \\ &= E\left[B_0 - b_0\right]^2 + E\left[\left(B_1 - b_1\right) X_{T+1}\right]^2 + E\left[U_{T+1}\right]^2 + E\left[2\left(B_0 - b_0\right)\left(B_1 - b_1\right) X_{T+1}\right] \\ &= V\left(b_0\right) + V\left(b_1\right) X_{T+1}^2 + V(U) + 2Cov\left(b_0, b_1\right) X_{T+1} \end{aligned}$$

assuming that *X* is constant in repeated sampling. Replacing the variances and the covariance with the expression for the sample estimators and rearrange we end up with the following expression:

$$\sigma_f^2 = E\left[Y_{T+1} - \hat{Y}_{T+1}\right]^2 = \sigma^2 \left[1 + \frac{1}{T} + \frac{\left(X_{T+1} - \bar{X}\right)}{\sum_{t=1}^{T} \left(X_t - \bar{X}\right)^2} \right] \qquad (4.12)$$

Observe that the forecast error variance is smallest when the future value of *X* equals the mean value of *X*. This formula is true if the future value of *X* is known. That is often not the case and hence the formula has to be elaborated accordingly. One way to deal with the uncertainty is to impose a distribution for *X*, with a component of uncertainty. That is, assume that

$$X_{T+1}^{*} = X_{T+1} + \varepsilon_{T+1}, \quad \varepsilon \sim N\left(0, \sigma_{\varepsilon}^2\right)$$

With this assumption we may form an expression for the error variance that takes the extra variation from the uncertainty into account:

$$\sigma_f^2 = \sigma^2 \left[1 + \frac{1}{T} + \frac{\left(X_{T+1} - \overline{X}\right)}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2} + \frac{\overline{X}_{T+1}\sigma_\varepsilon^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2} \right] + B_1^2\sigma_\varepsilon^2 \qquad (4.13)$$

The important point to notice here is that this variance is impossible to estimate unless we know the exact value of the variance for the uncertainty. That is of course not possible. Furthermore, the expression involves the population parameter multiplied with the variance of the uncertainty. Hence, in practice (4.12) is often use, but one should hold in mind that it most likely is an understatement of the true forecast error variance.

Taking the square root of the variance in (4.12) or (4.13) gives us the standard error of the forecast. With this standard error it is possible to calculate confidence interval around the predicted values using the usual formula for a confidence interval, that is:

Confidence interval of a forecast

$$\hat{Y}_{T+1} \pm t_c \times \sigma_f$$

5. Model measures

In the previous chapters we have developed the basics of the simple regression model, describing how to estimate the population parameters using sample information and how to perform inference on the population. But so far we do not know how well the model describes the data. The two most popular measures for model fit are the so called coefficient of determination and the adjusted coefficient of determination.

5.1 The coefficient of determination (R^2)

In the simple regression model we explain the variation of one variable with help of another. We can do that because they are correlated. Had they not been correlated there would be no explanatory power in our *X* variable. In regression analysis the correlation coefficient and the coefficient of determination are very much related, but their interpretation differs slightly. Furthermore, the correlation coefficient can only be used between pairs of variables, while the coefficient of determination can connect a group of variable with the dependent variable.

In general the correlation coefficient offers no information about the causal relationship between two variables. But the attempt of this chapter is to put the correlation coefficient in a context of the regression model and show under what conditions it is appropriate to interpret the correlation coefficient as a measure of strength of a causal relationship.

The coefficient of determination tries to decompose the average deviation from the mean into an explained part and an unexplained part. It is therefore natural to start the derivation of the measure from the deviation from mean expression and then introduce the predicted value that comes from the regression model. That is, for a single individual we have:

$$Y_i - \overline{Y} = Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i = \underbrace{\left(\hat{Y}_i - \overline{Y}\right)}_{\text{Explained}} + \underbrace{\left(Y_i - \hat{Y}_i\right)}_{\text{Unexplained}} \tag{5.1}$$

We have to remember that we try to explain the deviation from the mean value of *Y*, using the regression model. Hence, the difference between the expected value $\left(\hat{Y}_i\right)$ and the mean value $\left(\overline{Y}\right)$ will therefore be denoted as the explained part of the mean difference. The remaining part will therefore be denoted the unexplained part. With this simple trick we decomposed the simple mean difference for a single observation. We must now transform (5.1) into an expression that is valid for the whole sample, that is for all observations. We do that by squaring and summing over all *n* observations:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\left(\hat{Y}_i - \overline{Y}\right) - \left(Y_i - \hat{Y}_i\right)\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{Y}_i - \overline{Y}\right)^2 + \left(Y_i - \hat{Y}_i\right)^2 - 2\left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right)\right]$$

It is possible to show that the sum of the last expression on the right hand side equals zero. With that knowledge we may write:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}_{TSS}=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}_{ESS}+\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{RSS} \tag{5.2}$$

With these manipulations we end up with three different components. On the left hand side we have the total sum of squares (TSS) which represents the total variation of the model. On the right hand side we first have the Explained Sum of Squares (ESS) and the second component on the right hand side represents the unexplained variation and is called thee Residual Sum of Squares (RSS).

Caution: be careful when using different text books. The notation is not consistent in the literature so it is always important to make sure that you know what ESS and RSS stands for.

The identity we have found may now be expressed as:

$$TSS = ESS + RSS$$

which may be rewritten in the following way:

$$\frac{TSS}{TSS}=\frac{ESS}{TSS}+\frac{RSS}{TSS}=1$$

Hence, by dividing by the total variation on both sides we may express the explained and unexplained variation as shares of the total variation, and since the right hand side sum to one, the two shares can be expressed in percentage form. We have

$\frac{ESS}{TSS}$ = the share of the total variation that is explained by the model

$\frac{RSS}{TSS}$ = the share of the total variation that is unexplained by the model

The coefficient of determination

The percent of variation in the dependent variable associated with or explained by variation in the independent variable in the regression equation:

$$R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}, \qquad 0\le R^2\le 1$$

Example 5.1

Assume that a simple linear regression model estimated an R^2 equal to 0.65. That would imply that 65 percent of the total variation around the mean value of *Y* is explained by the variable *X* included in the model.

In the simple regression model there is a nice relationship among the measures of sample correlation coefficient, the OLS estimator of the slope coefficient, and the coefficient of determination. To see this we may rewrite the explained sum of squares in the following way:

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\left(b_0 + b_1 X_i\right) - \left(b_0 + b_1 \overline{X}\right)\right)^2 = \sum_{i=1}^{n}\left(b_1 X_i - b_1 \overline{X}\right)^2 = b_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Using this transformation we may re-express the coefficient of determination:

$$R^2 = \frac{ESS}{TSS} = \frac{b_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \left(b_1 \frac{S_X}{S_Y}\right)^2 \tag{5.3}$$

where S_X and S_Y represents the sample standard deviation for *X* and *Y* respectively. Furthermore we can establish a relation between the OLS slope estimator and the correlation coefficient between *X* and *Y*.

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X^2} \times \frac{S_Y}{S_Y} = \frac{S_{XY}}{S_X S_Y} \times \frac{S_Y}{S_X} = r\frac{S_Y}{S_X} \tag{5.4}$$

where S_{XY} represents the sample covariance between *X* and *Y*, and *r* the sample correlation coefficient for *X* and *Y*. Hence, substituting (5.4) into (5.3) shows the relation between the sample correlation coefficient and the coefficient of determination.

$$R^2 = \left(b_1 \frac{S_X}{S_Y}\right)^2 = \left(r \times \frac{S_X}{S_Y} \times \frac{S_Y}{S_X}\right)^2 = (r)^2 \tag{5.5}$$

Hence, in the simple regression case the square root of the coefficient of determination is the sample correlation coefficient:

$$r = \sqrt{R^2} \tag{5.6}$$

This means that the smaller the correlation between *X* and *Y*, the smaller is the explained share of the variation by the model, which is the same as to say that the larger is the unexplained share of the variation. That is, the more disperse the sample points are from the regression line the smaller is the correlation and the coefficient of determination. This leads to an important conclusion about the importance of the coefficient of determination:

R^2 and the significance of the OLS estimators

An increased variation in *Y*, with an unchanged variation in *X*, will directly reduce the size of the coefficient of determination. But it will not have any effect on the significance of the parameter estimate of the regression model.

From (5.3)-(5.6) it is clear that an increased variation in *Y* will reduce the size of the coefficient of determination of the regression model. However, when the variation in *Y* increases, so will the covariance between *Y* and *X* which will increase the value of the parameter estimate. It is therefore not obvious that the significance of the parameter will be unchanged. By creating the t-ratio we can see that:

$$t = \frac{b_1}{se(b_1)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Bigg/ \sqrt{\frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n-2}}}$$

where S represents the standard deviation of the residual. The expression for the standard error of the OLS estimator was derived in the previous chapter. Now, lets us see what happens with the *t*-value if we increase the variation of *Y* with a constant *c*.

$$\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(cY_i-c\overline{Y}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(cY_i-c\hat{Y}_i\right)^2}{n-2}}}=\frac{\sum_{i=1}^{n}c\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\frac{c^2\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-2}}}=\frac{c}{c}\times\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-2}}}=\frac{b_1}{se(b_1)}=t$$

Hence, increasing the variation of Y with a constant c, has no effect what so over on the t-value. We should therefore draw the conclusion that the coefficient of determination is just a measure of linear strength of the model and nothing else. As an applied researcher it is far more interesting and important to analyze the significance of the parameters in the model which is related to the t-values.

5.2 The adjusted coefficient of determination (Adjusted R^2)

The coefficient of determination can be used for describing the linear strength of a regression model. But its size is dependent on the degrees of freedom. Therefore it is not meaningful to compare R^2 between two different models with different degrees of freedom. A solution to this problem is to control for the degrees of freedom and adjust the coefficient of determination accordingly. That could be done in the following way:

$$\overline{R}^2=1-\frac{RSS/(n-2)}{TSS/(n-1)}=1-\frac{Var(e)}{Var(Y)} \qquad (5.7)$$

where $\overline{R}^2$ denotes the adjusted coefficient of determination. The adjusted coefficient of determination can also be expressed as a function of the unadjusted coefficient of determination in the following way:

$$\overline{R}^2=1-\frac{RSS/(n-2)}{TSS/(n-1)}=1-\frac{RSS}{TSS}\times\frac{n-1}{n-2}=1-\frac{TSS-ESS}{TSS}\times\frac{n-1}{n-2}=1-\left(1-R^2\right)\frac{n-1}{n-2}$$

It turns out that the adjusted R^2 is always lower than the unadjusted R^2, and when the number of observations increases they converge to each other. Using equation (5.7) we see that the adjusted coefficient of determination is a function of the variance of Y as well as the variance of the residual. Rearranging (5.7) we receive that

$$S^2=Var(e)=\left(1-\overline{R}^2\right)S_Y^2$$

As can be seen, the larger the adjusted R^2, the smaller become the residual variance. Another interesting feature of the adjusted R^2 is that it can be negative, an event impossible for the unadjusted R^2. That is especially likely to happen when the number of observations are low, and the unadjusted R^2 is small, let's say around 0.06.

Another important point to understand is that the coefficient of determination can only be compared between models when the dependent variable is the same. If you have Y in one model and ln(Y) in another, the dependent variable is transformed and should not be treated as the same. It is for that reason not meaningful to compare the adjusted or unadjusted R^2 between these two models.

5.3 The analysis of variance table (ANOVA)

Almost all econometric software generates an ANOVA table together with the regression results. An ANOVA table includes and summarizes the sum of squares calculated above:

Table 5.1 The ANOVA Table

Source of Variation	Degrees of freedom	Sum of Squares	Mean Squares
Explained	1	*ESS*	*ESS*/1
Unexplained	*n*-2	*RSS*	*RSS*/(*n*-2)F=*MSE*=S^2
Total	*n*-1	*TSS*	

The decomposition of the sample variation in *Y* can be used as an alternative approach of performing test within the regression model. We will look at two examples that work for the simple regression model. In the multiple regression case we have an even more important use, which will be described in chapter 6, and is related to simultaneous test on sub sets of parameters.

Assume that we are working with the following model:

$$Y_i = B_0 + B_1 X_i + U_i$$

Using a random sample we calculated the components of the ANOVA table want to perform a test for the following hypothesis:

$$H_0 : B_1 = 0$$
$$H_1 : B_1 \neq 0$$

Remember that the ANOVA table contains information about the explained and unexplained variation. Hence if the explained part increases sufficiently by including *X*, we would be able to say that the alternative hypothesis is true. One way to measure this increase would be to use the following ratio:

$$F = \frac{\text{Additional variance explained by } X}{\text{Unexplained variance}} \qquad (5.8)$$

In the numerator of equation (5.8) we have the change in explained sum of squares divide by the degrees of freedom that come from including an additional variable in the regression model. Since this is a simple regression model the explained part goes from zero since no other variables are included and therefore the degrees of freedom equals one. Hence, the expression in the numerator is therefore simply the ESS. In the denominator we have the variance of the residual. It turns out that the ratio of the two components has a known distribution that is tractable to work with. That is:

$$F = \frac{ESS/1}{RSS/(n-2)} \sim F_{(1,n-2)} \qquad (5.9)$$

Hence, we have a test function that is F-distributed with 1 and *n*-2 degrees of freedom.

Example 5.2

Assume that we have a sample of 145 observations and that we would like to know if the random variable *X* has any effect on the dependent variable *Y*. In order to answer this question we form a simple regression model, and form the following hypothesis: $H_0 : B_1 = 0$ vs. $H_1 : B_1 \neq 0$. Use the following information to perform the test:

$$ESS = 51190, \quad RSS = 5232$$

In order to carry out the test, we form the test function and calculate the corresponding test value. Using (5.9) we receive:

$$F = \frac{51190/1}{5232/(145-2)} = 1399.1$$

With a significance level of 5 percent we receive the following upper critical value, $F_{0.025}(1{,}143) = 5.13$, which is very much lower than the test value. Hence we can reject the null hypothesis and conclude that *X* has a significant effect on *Y*.

When using the ANOVA table to perform a test on the parameters of the model we call this the test of over all significance. In the simple regression model case it involves just one single parameter, but in the multiple variable case the test consider the joint hypothesis that all the included variables has a joint effect that is zero. We will speak more about this in the next chapter.

In the simple regression case the F-test corresponds to the simple t-test related to the slope coefficient. But how are these two test functions connected. To see this, we may rewrite the F-test in the following way:

$$F=\frac{ESS/1}{RSS/(n-2)}=\frac{b_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{S^2}=\frac{b_1^2}{S^2\Big/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}=\frac{b_1^2}{V(b_1)}=\left(\frac{b_1}{se(b_1)}\right)^2=t^2$$

The F-statistic in this case is nothing more than the square of the t-statistic of the regression coefficient. Hence, the outcomes of the two procedures are always consistent.

6. The multiple regression model

From now on the discussion will concern multiple regression analysis. Hence, the analysis will be assumed to include all relevant variables that explain the variation in the dependent variable, which almost always includes several explanatory variables. That has consequences on the interpretation of the estimated parameters, and violations to this condition will have consequences that will be discussed in chapter 7. This chapter will focus on the differences between the simple and the multiple-regression model and extend the concepts from the previous chapters.

6.1 Partial marginal effects

For notational simplicity we will use two explanatory variables to represent the multiple-regression model. The population regression function would now be expressed in the following way:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{6.1}$$

By including another variable in the model we control for additional variation that is attributed to that variable. Hence coefficient B_1 represents the unique effect that comes from X_1, controlling for X_2, which means that, any common variation between X_1 and X_2 will be excluded. We are talking about the **partial regression coefficient**.

Example 6.1
Assume that you would like to predict the value (sales price) of a Volvo S40 T4 and you have access to a data set including the following variables: sale price (P) the age of the car (A) and the number of kilometers the car has gone (K). You set up the following regression model:

$$P = B_0 + B_1 A + B_2 K + U \tag{6.2}$$

The model offers the following two marginal effects:

$$\frac{\partial P}{\partial A} = B_1 \tag{6.3}$$

$$\frac{\partial P}{\partial K} = B_2 \tag{6.4}$$

The first marginal effect (6.3) represents the effect from a unit change in the age of the car on the conditional expected value of sales prices. When the age of the car increase by one year, the mean sales price increase by B_1 Euros when controlling for number kilometers. It is reasonable to believe that the age of the car is correlated with the number of kilometers the car has gone. That means that some of the variations in the two variables are common in explaining the variation in the sales price. That common variation is excluded from the estimated coefficients. The partial effect that we seek is therefore the unique effect that comes from the aging of the car.

Accordingly, the second marginal effect (6.4) represents the unique effect that each kilometer has on the sales price of the car, controlling for the age of the car. The way the model is specified here, imply that the unique effect on the sales price from each kilometer is the same whether the car is new or if it is 10 years old, which means that the marginal effects are independent of the level of *A* and *K*. If this is implausible, one could adjust for it.

One way to extend the model and control for additional variation would be to include squared terms as well as cross products. The extended model would then be:

$$P = B_0 + B_1 A + B_2 A^2 + B_3 K + B_4 K^2 + B_5 A \times K + U \tag{6.5}$$

Extending the model in this way would results in the following two marginal effects:

$$\frac{\partial P}{\partial A} = B_1 + 2B_2 A + B_5 K \tag{6.6}$$

$$\frac{\partial P}{\partial K} = B_3 + 2B_4 K + B_5 A \tag{6.7}$$

Equation (6.6) is the marginal effect on sales price from a unit increase in age. It is a function of how old the car is and how many kilometer the car has gone. In order to receive a specific vale for the marginal effect we need to specify values for *A* and *K*. Most often those values would be mean values of *A* and *K*, unless other specific values are of particular interest. The marginal effects given by (6.6) and (6.7) consist of three parameter estimates, which individually can be interpreted.

Focusing on (6.6) the first parameter estimate is B_1. It should be regarded as an intercept, and as such has limited interest. Strictly speaking it represents the marginal effect, when *A* and *K* both are zero, which would be when the car was new.

The second parameter is B_2 that accounts for any non-linear relation between *A* and *P*. To include a squared term is therefore a way to test if the relation is non-linear. If the estimated coefficient is significantly different from zero we should conclude that non-linearity is present and controlling for it would be necessary. Failure to control for it, would lead to a biased marginal effect since it would be assumed to be constant, when it in fact vary with the level of A.

The third parameter B_5 controls for any synergy effect that could possible exist between the two explanatory variables included. It is not obvious that such effect would exist in the Volvo S40 example. In other areas of economics the effect is more common. For instance in the US wage equation literature: being black and being a woman are usually two factors that have negative effects on the wage rate. Furthermore, being a black woman is a combined effect that further reduces the wage rate. This would be an example of a negative synergy effect.

6.2 Estimation of partial regression coefficients

The mathematics behind the estimation of the OLS estimators in the multiple regression case is very similar to the simple model, and the idea is the same. But the formulas for the sample estimators are slightly different. The sample estimators for model (6.2) are given by the following expressions:

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 \tag{6.8}$$

$$b_1 = \frac{S_{Y1} - r_{12}r_{Y2}S_Y S_1}{S_1^2(1-r_{12}^2)} \tag{6.9}$$

$$b_2 = \frac{S_{Y2} - r_{12}r_{Y1}S_Y S_2}{S_2^2(1-r_{12}^2)} \tag{6.10}$$

where S_{Y1} is the sample covariance between Y and X_1, r_{12} is the sample correlation between X_1 and X_2, r_{Y2} is the sample correlation between Y and X_2, S_Y is the sample standard deviation for Y, and S_1 is the sample standard deviation for X_1. Observe the similarity between the sample estimators of the multiple-regression model and the simple regression model. The intercept is just an extension of the estimator for the simple regression model, incorporating the additional variable. The two partial regression slope coefficients are slightly more involved but possess an interesting property. In case of (6.9) we have that

$$b_1 = \frac{S_{Y1} - r_{12}r_{Y2}S_Y S_1}{S_1^2(1-r_{12}^2)} = \frac{S_{Y1}}{S_1^2} \qquad \text{if } r_{12} = 0$$

That is, if the correlation between the two explanatory variables is zero, the multiple regression coefficients coincide with the sample estimators of the simple regression model. However, if the correlation between X_1 and X_2 equals one (or minus one), the estimators are not defined, since that would lead to a division by zero, which is meaningless. High correlation between explanatory variables is referred to as a collinearity problem and will be discussed further in chapter 11. Equation (6.8)-(6.10) can be generalized further to include more parameters. When doing that all par wise correlations coefficients are then included in the sample estimators and in order for them to coincide with the simple model, they all have to be zero.

The measure of fit in the multiple regression case follows the same definition as for the simple regression model, with the exception that the coefficient of determination no longer is the square of the simple correlation coefficient, but instead something that is called the **multiple-correlation coefficient**.

In multiple regression analysis, we have a set of variables X_1, X_2, ... that is used to explain the variability of the dependent variable *Y*. The multivariate counterpart of the coefficient of determination R^2 is the coefficient of multiple determination. The square root of the coefficient of multiple determination is the coefficient of **multiple correlation**, *R*, sometimes just called the multiple *R*. The multiple *R* can only take positive values as appose to simple correlation coefficient that can take both negative and positive values. In practice this statistics has very little importance, even though it is reported in output generated by softwares such as Excel.

6.3 The joint hypothesis test

An important application of the multiple regression analysis is the possibility to test several parameters simultaneously. Assume the following multiple-regression model:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + U \qquad (6.11)$$

Using this model we may test the following hypothesis:

a) $H_0 : B_1 = 0$ vs. $H_1 : B_1 \neq 0$
b) $H_0 : B_1 = B_2 = 0$ vs. H_1: H_0 not true
c) $H_0 : B_1 = B_2 = B_3 = 0$ vs. H_1: H_0 not true

The first hypothesis concerns a single parameter test, and is carried out in the same way here as was done in the simple regression model. We will therefore not go through these steps again but instead focus on the simultaneous tests given by hypothesis *b* and *c*.

6.3.1 Testing a subset of coefficients

The hypothesis given by (*b*) represents the case of testing a subset of coefficients, in a regression model that contains several (more than two) explanatory variables. In this example we choose to test B_1 and B_2 but it could of course be any other combination of pairs of coefficients included in the model. Let us start by rephrasing the hypothesis, with the emphasis on the alternative hypothesis:

$$H_0 : B_1 = B_2 = 0$$
$$H_1 : B_1 \neq 0 \; and / or \; B_2 \neq 0$$

It is often believed that in order to reject the null hypothesis, both (all) coefficients need to be different from zero. That is just wrong. It is important to understand that the complement of the null hypothesis in this situation is represented by the case where at least one of the coefficients is different from zero.

Whenever working with test of several parameters simultaneously we cannot use the standard t-test, but instead we should be using an F-test. An F-test is based on a test statistic that follows the F-distribution. We would like to know if the model that we stated is equivalent to the null hypothesis, or if the alternative hypothesis is a significant improvement of the fit. So, we are basically testing two specifications against each other, which are given by:

Model according to the null hypothesis: $Y = B_0 + B_3 X_3 + U$ (6.12)

Model according to the alternative hypothesis: $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + U$ (6.13)

A way to compare these two models is to see how different their *RSS* (Residual Sum of Squares) are from each other. We know that the better fit a model has, the smaller is the *RSS* of the model. When looking at specification (6.12) you should think of it as a restricted version of the full model given by (6.13) since two of the parameters are forced to zero. In (6.13) on the other hand, the two parameters are free to take any value the data allows them to take. Hence, the two specifications generate a Restricted *RSS* (RSS_R) received from (6.12) and an Unrestricted *RSS* (RSS_U) received from (6.13). In practices this means that you have to run each model separately using the same data set and collect *RSS*-values from each regression and then calculate the test value.

The test value can be received from the test statistic (test function) given by the following formula:

$$F = \frac{(RSS_R - RSS_U)/df_1}{RSS_U / df_2} \sim F_{(\alpha, df_1, df_2)} \tag{6.14}$$

where df_1 and df_2 refers to the degrees of freedom for the numerator and denominator respectively. The degrees of freedom for the numerator is simply the difference between the degrees of freedom of the two Residual Sum of Squares. Hence, $df_1 = (n\text{-}k_1) – (n\text{-}k_2) = k_2\text{-}k_1$. k_1 is the number of parameters in the restricted model, and k_2 is the number of parameters in the unrestricted model. In this case we have that $k_2\text{-}k_1\text{=}2$.

When there is very little difference in fit between the two models the difference given in the numerator will be very small and the F-value will be close to zero. However, if the fit differ extensively, the F-value will be large. Since the test statistic given by (6.14) has a know distribution (if the null hypothesis is true) we will be able to say when the difference is sufficiently large to say that the null hypothesis should be rejected.

Example 6.2
Consider the two specifications given by (6.12) and (6.13), and assume that we have a sample of 1000 observations. Assume further that we would like to test the joint hypothesis discussed above. Running the two specifications on our sample we received the following information given in Table 6.1.

Table 6.1 Summary results from the two regressions

The Restricted Model	The Unrestricted Model
k_1 = 2	k_2 = 4
RSS = 17632	*RSS* = 9324

Using the information in Table 6.1 we may calculate the test value for our test.

$$F = \frac{(RSS_R - RSS_U)/df_1}{RSS_U / df_2} = \frac{(17632 - 9324)/(4-2)}{9324/(1000-4)} = \frac{4154}{9.36144} = 443.73$$

The calculated test value has to be compared with a critical value. In order to find a critical value we need to specify a significance level. We choose the standard level of 5 percentage and find the following value in the table: $F_C = 4.61$.

Observe that the hypothesis that we are dealing with here is one sided since the restricted *RSS* never can be lower than the unrestricted *RSS*. Comparing the critical value with the test value we see that the test value is much larger, which means that we can reject the null hypothesis. That is, the parameters involved in the test have a simultaneous effect on the dependent variable.

6.3.2 Testing the regression equation

This test is often referred to as the test of the over all significance and by performing the test we ask if the included variables has a simultaneous effect on the dependent variable. Alternatively, we ask if the population coefficients (excluding the intercept) are simultaneously equal to zero, or at least one of them are different from zero.

In order to test this hypothesis, we compare the following model specifications against each other:

Model according to the null hypothesis: $Y = B_0 + U$ (6.15)

Model according to the alternative hypothesis: $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + U$ (6.16)

The test function that should be used for this test is the same in structure as before, but with some important differences, that makes it sufficient to estimate just one regression for the full model instead of one for each specification. To see this we can rewrite the RSS_R in the following way:

$$RSS_R = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - b_0\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = TSS_U$$

Hence the test function can be expressed in sums of squares that could be found in the ANOVA table of the unrestricted model. The test function therefore becomes:

$$F = \frac{\left(RSS_R - RSS_U\right)/\,df_1}{RSS_U\,/\,df_2} = \frac{\left(TSS_U - RSS_U\right)/\,df_1}{RSS_U\,/\,df_2} = \frac{ESS/(k-1)}{RSS/(n-k)}$$

Example 6.3

Assume that we have access to a sample of 1000 observations and that we would like to estimate the parameters in (6.16), and test the over all significance of the model. Running the regression using our sample we received the following ANOVA table:

Table 6.2 ANOVA table

Variation	Degrees of freedom	Sum of squares	Mean squares
Explained	3	4183	1394.33
Residual	996	1418	1.424
Total	999	5602	

Using the information from Table 6.2 we can calculate the test value:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{1394.33}{1.424} = 979.37$$

This is a very large test value. We can therefore conclude that the included parameters explains a significant part of the variation of the dependent variable.

7. Specification

The formulation of a satisfactory econometric model is very important if we are to draw any conclusion from it. Sometimes the underlying theory of the model gives us guidelines on how it should be specified, but in other cases we have to rely on statistical tests. In this chapter we will discuss the most important issues related to the formulation and specification of an econometric model.

7.1 Choosing the functional form

To choose a correct functional form is very important since it has implication on the interpretation of the parameters that are estimated. When formulating the model we need to know how the coefficients are to be interpreted, and how the marginal effect and elasticity looks like. Below we will go through the most basic functional forms and describe when they can be used.

7.1.1 The linear specification

When talking about a linear specification we have to remember that all models that we are talking about in this text are linear in their parameters. Any deviation from linearity will therefore only be related to the relation between the variables. The linear specification is appropriate when *Y* and *X* has a linear relation. The econometric model would then be expressed in this way:

$$Y = B_0 + B_1 X + U \tag{7.1}$$

For simplicity reasons we express the model as the simple regression model. The interpretation of the slope coefficient coincide with the marginal effect, which is

$$\frac{dY}{dX} = B_1 \tag{7.2}$$

This means that when *X* increases by 1 unit, *Y* will change by B_1 units. Since it is expressed in units it is important to remember in what unit form the data is organized. If *Y* represents the yearly disposable income expressed in thousands of Euro and *X* represents age given in years, we have to understand that a unit change in *X* represents a year and the corresponding effect in the yearly disposable income is in thousands of Euros.

Since the unit change usually is dependendent on the level of the dependent variable the marginal effect has some limitation. That is, the effect of a unit change in *X* may be different whether the level of *X* is 10 or if it is 10,000. Therefore economists usually prefer to analyze the elasticities instead of the marginal effect, and in the linear model the elasticity is given by:

$$e = \frac{dY}{dX}\frac{X}{Y} = B_1 \frac{X}{Y} \tag{7.3}$$

which usually is expressed using mean values of *X* and *Y*. The elasticity denoted by *e* is not expressed in terms of units, but instead expressed in relative terms. A 1 percent increase in *X* will result in *e* percent change in *Y*.

Example 7.1
Calculate the marginal effect and the elasticity using the regression results in Table 7.1 received from a sample of data using model (7.1).

Table 7.1 Regression results from a model with a linear specification

	Coefficients	Standard error		Mean values
Intercept	1.066	4.944	Mean of *X*	5.5
X	8.557	0.796	Mean of *Y*	48.1

The marginal effect from a variable *X* on *Y* using a linear specification is received directly from the parameter estimate. In this example we receive

$$\frac{dY}{dX} = 8.6$$

Hence, when *X* increase by 1 unit, *Y* increases by 8.6 units. The more important measure of elasticity is here given by:

$$e = \frac{dY}{dX}\frac{\overline{X}}{\overline{Y}} = 8.6 \times \frac{5.5}{48.1} = 0.983$$

That is, when *X* increases by 1 percent, the dependent variable *Y* increases by 0.98 percent.

7.1.2 The log-linear specification

In the log linear specification the relationship between *X* and *Y* is no longer linear and is written as

$$\ln Y = B_0 + B_1 X + U \tag{7.4}$$

Several factors could motivate this specification. This specification is widely used in the human capital literature where economic theory suggests that earnings should be in logarithmic form when estimating the return to education on earnings. This could be motivated in the following way: assume that the rate of return to an extra year of education is denoted by *r*. Given an initial period earnings, w_0, the first year of schooling would generate an earnings equal to $w_1 = (1+r)w_0$. After *s* years of schooling we would have an earnings equal to: $w_s = (1+r)^s w_0$. Taking the logarithms of this we receive

$$\ln(w_s) = s\ln(1+r) + \ln(w_0) = B_0 + B_1 s \tag{7.5}$$

which is a log-linear relationship between years of schooling and earnings. With a similar motivation we could include several other variables, such as age and years of work experience which the theory states are important factors for the earnings generation. By including an error term we form a statistical model in the form of (7.4). How do we interpret the slope parameter? It is important to remember that we are primarily intereted in the effect on earnings not the logarithm of the earnings. Hence, it is not possible or meaningful to say that a unit increase in *s* will result in a unit change in the logarithm of earnings. Taking the derivative of earnings with respect to schooling gives us the following expression:

$$\frac{d\ln(w_s)}{ds} = \frac{1}{w_s}\frac{dw_s}{ds} = \frac{dw_s / w_s}{ds} = B_1 \tag{7.6}$$

Expression (7.6) shows us that the slope coefficient should be interpreted as the relative change in earnings as a ratio of the absolute change in schoolings. In other words, if schooling increase by one year, earnings will change by $B_1 \times 100$ percent.

Using (7.6) we see that the marginal effect is given by:

$$\frac{dw_s}{ds} = B_1 \times w_s \tag{7.7}$$

Hence, the marginal effect is an increasing function of earnings itself. That is, if schooling increases by one year, earnings will change by $B_1 \times w_s$ units. Hence the response on the dependent variable will change in terms of unit, but is constant in relative terms.

Using (7.7) we can derive the earnings elasticity with respect to years of schooling:

$$e = \frac{dw_s}{ds}\frac{s}{w_s} = B_1 w_s \frac{s}{w_s} = B_1 \times s$$

The earnings elasticity is an increasing function of the number of years of schooling. Hence the longer you have studied the larger is the earnings elasticity.

7.1.3 The linear-log specification

In the linear-log model it is the explanatory variable that is expressed and transformed using the logarithmic transformation which appears as follows

$$Y = B_0 + B_1 \ln X + U \tag{7.8}$$

Taking the derivative of *Y* with respect to *X* we receive:

$$\frac{dY}{dX} = \frac{1}{X} B_1 \qquad <=> \qquad \frac{dY}{dX / X} = B_1 \tag{7.9}$$

Hence, the parameter estimate of the slope coefficient is the absolute change in *Y* over the relative change in *X*, which is to say that if *X* increase by 1 percent, the dependent variable *Y* will change by B_1 units.

Using the expression for the coefficient we may write the elasticity as follows:

$$e = \frac{dY}{dX}\frac{X}{Y} = \frac{1}{X} B_1 \frac{X}{Y} = \frac{1}{Y} B_1$$

Hence the elasticity is a function of the dependent variable, and the larger the dependent variable is the smaller become the elasticity, everything else equal.

7.1.4 The log-log specification

The log-log specification is another important and commonly used specification that can be motivated by the economic model. The so called Cobb-Douglass functions are often used as production functions in economic theories. They are usually expressed as follows:

$$Q(L, K) = AL^{B_1} K^{B_2} \tag{7.10}$$

(7.10) is a commonly used production function that is a function of two variables; labor (*L*) and capital (*K*). This model is multiplicative and non linear in nature which makes it difficult to use. However, there is an easy way to make this model linear and that is by means of taking the logarithm of both sides. Doing that and adding an error term we receive:

$$\ln Q = \ln A + B_1 \ln L + B_2 \ln K + U = B_0 + B_1 \ln L + B_2 \ln K + U \tag{7.11}$$

Hence, the so called log-log specification requires that both left hand and right hand side of the equation are in logarithmic form, and that is what we have in (7.11). Furthermore, it is also linear in the parameters which make it easy to estimate statistically. How do we interpret these parameters? Let us focus on B_1 when answering this question. Take the derivative of *Q* with respect to *L* and receive:

$$\frac{\partial \ln Q}{\partial \ln L} = \frac{\partial Q / Q}{\partial L / L} = \frac{\partial Q}{\partial L}\frac{L}{Q} = B_1 \tag{7.12}$$

The coefficients of the log-log model are conveniently expressed as elasticities. So the elasticity and the coeffieints coincide. Remember that the elasticity is expressed in percentage and not in decimal form and hence should not be multiplied by 100.

The marginal effect of the log-log model can be received from (7.12) and equals:

$$\frac{\partial Q}{\partial L} = B_1 \frac{Q}{L} \tag{7.13}$$

which is a function of both *Q* and *L*. Hence the marginal effect is increasing in *Q* and decreasing in *L*, everything else equal.

7.2 Omission of a relevant variable

In chapter 3 we described how the error term could be seen as collection of everything that is not accounted for by observable variables included in the model. We should also remember that the first assumption related to the regression model concerns the fact that all that is relevant should be included in the model. What are the consequences of not including everything that is relevant in the model?

In order to answer that question we need to know the meaning of the word relevance. Unfortunately it has several meaning and we usually make the distinction between statistical and economic relevance. Statistical relevance refers to whether the coefficient is significantly different from zero or not. That is, if we are able to reject the null hypothesis. If we are unable to reject the null hypothesis we say that the variable has no statistical relevance.

The economic relevance is related to the underlying theory that the model is based on. Variables are included in the model because the economic theory says they should be. That some of the variables are not significantly different from zero is not a criterion for exclusion. It is the economic relevance that makes the omission of a relevant variable problematic. To see this consider the following two specifications:

The correct economic model: $Y = B_0 + B_1 X_1 + B_2 X_2 + U$ (7.14)

The estimated model: $Y = b_0 + b_1 X_1 + e$ (7.15)

From chapter 3 we know that the sample estimator for the slope coefficient in the simple regression model is given by:

$$b_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)Y_i}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(B_0 + B_1 X_{1i} + B_2 X_{2i} + U_i)}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} \tag{7.16}$$

which may be rewritten as

$$b_1 = B_0 \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} + B_1 \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)X_{1i}}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} + B_2 \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)X_{2i}}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} + \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)U_i}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} \tag{7.17}$$

Simplify and take the expectation of the estimator:

$$E[b_1] = B_1 + B_2 \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)X_{2i}}{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2} = B_1 + B_2 \frac{Cov(X_1, X_2)}{Var(X_1)} \tag{7.18}$$

Hence, the estimator is not unbiased any more. The expected value of the estimator is a function of the true population parameter of B_1 and the true population parameter B_2 times a weight that persist even if the number of observations goes to infinity. Failure to include all relevant variables therefore makes the coefficients of the included variables biased and inconsistent. However, if the excluded variable is statistically independent of the included variable, that is if the covariance between X_1 and X_2 is zero, exclusion will not be a problem, since the second component of (7.18) will equal zero, and the estimator will be unbiased. If the model includes several variables and one relevant variable is excluded, the bias will affect all the coefficients as long as the corresponding variables are correlated with the excluded variable.

A common example of this kind of bias appears in the human capital literature when they try to estimate the return to education on earnings without including a variable for scholastic ability. The problem is common since most data set does not include such information and that scholastic ability is correlated with the number of years of schooling as well as earnings. Since scholastic ability is believed to be positively correlated with schooling as well as with the earnings, the rates of returns to education are usually overestimated, due to the second component in (7.18).

7.3 Inclusion of an irrelevant variable

Another situation that often appears is associated with adding variables to the equation that are economically irrelevant. The researcher might be keen on avoiding the problem of excluding any relevant variables, and therefore include variables on the basis of their statistical relevance. Some of the included variables could then be irrelevant economically, which have consequences on the estimated coefficients. The important question to ask is what those consequences are. To see what happens when including economically irrelevant variables we start by defining two equations:

The correct economic model: $Y = B_0 + B_1 X_1 + U$ (7.19)

The estimated model: $Y = b_0 + b_1 X_1 + b_2 X_2 + e$ (7.20)

The estimated model (7.20) includes two variables, and X_2 is assumed to be economically irrelevant, which means that its coefficient is of minor interest. The OLS estimator of the coefficient for the other variable is given by:

$$b_1 = \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)Y_i - \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)Y_i}{\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 - \left(\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)\right)^2} \tag{7.21}$$

Substitute the (7.19) for Y and take the expectation to obtain

$$E[b_1] = \frac{B_1\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 - B_1\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{1i}-\bar{X}_1)}{\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 - \left(\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)\right)^2} = B_1$$

Hence, the OLS estimator is still unbiased. However, the standard error of the estimator is larger when including extra irrelevant variables, compared to the model where only the relevant variables are included, since more variation is added to the model. Therefore, the price of including irrelevant variables is in efficiency and the estimator is no longer BLUE. On the other hand loss in efficiency is less harmful compared to biased and inconsistent estimates. Therefore, when one is unsure about a model specification, one is better off including too many variables, than too few. This is sometimes called **kitchen-sink regressions**.

7.4 Measurement errors

Until now we have assumed that all variables, dependent as well as independent, have been measured without any errors. That is seldom the case and therefore it is important to understand the consequences it has on the OLS estimator. We are going to consider three cases: measurement error in *Y* only, measurement error in *X* only, and measurement error in both *X* and Y.

In order to analyze the consequences of the first case we have to assume a structure of the error. We assume that the measurement error is random and defined in the following way:

$$Y^* = Y + \varepsilon \qquad (7.22)$$

where *Y** represent the observed variable, *Y* the true, and *ε* the random measurement error that is independent of *Y*, with a mean equal to zero and a fixed variance σ_ε^2 . Assume the following population model and substitute (7.22) with *Y*:

$$Y = B_0 + B_1 X + U$$
$$Y^* - \varepsilon = B_0 + B_1 X + U$$

$$Y^* = B_0 + B_1 X + (U + \varepsilon) = B_0 + B_1 X + U^* \qquad (7.23)$$

The new error term *U** would still be uncorrelated with the independent variable *X*, so the sample estimators would still be consistent and unbiased. That is, we have

$$Cov(X, U^*) = Cov(X, \varepsilon + U) = \underbrace{Cov(X, \varepsilon)}_{=0} + \underbrace{Cov(X, U)}_{=0} = 0$$

However, the new error would have a variance that are larger than otherwise, that is, $V(\varepsilon + U) = \sigma_U^2 + \sigma_\varepsilon^2$. Remember that the measurement error is random which imply that the population error term is uncorrelated with the measurement error. Hence the two variances only add to a larger total variance, which affects the standard errors of the estimates as well. The conclusion is that random measurement errors in the dependent variable do not matter much in practice.

In the second case the measurement error is attached to the independent variable, still under the assumption that the error is random. Assume that the observed variable is defined in the following way:

$$X^* = X + \varepsilon \qquad (7.24)$$

with an error component that is independent of *X*, has a mean zero and a fixed variance, σ_ε^2. When the observed explanatory variable is defined in this way the population regression equation is affected in the following way. The model we would like to study is defined as

$$Y = B_0 + B_1 X + U$$

but we only observe *X**, which implies that the model become

$$Y = B_0 + B_1 (X^* - \varepsilon) + U$$
$$Y = B_0 + B_1 X^* + (U - B_1 \varepsilon) = B_0 + B_1 X^* + U^* \qquad (7.25)$$

The mean value of the new error term is still zero, and the variance is some what inflated compared to the case with no measurement error. That is, $V(U^*) = V(U - B_1\varepsilon) = \sigma_U^2 + B_1^2 \sigma_\varepsilon^2$. Unfortunatelly the new error term is no longer uncorrelated with the explanatory variable. The measurement error creates a correlation that is different from zero, that bias the OLS estimators. That is

$$Cov(X^*, U^*) = Cov(X + \varepsilon, U - B_1\varepsilon) = \underbrace{Cov(X, U)}_{=0} - \underbrace{Cov(X, B_1\varepsilon)}_{=0} + \underbrace{Cov(\varepsilon, U)}_{=0} - Cov(\varepsilon, B_1\varepsilon) = -B_1^2 \sigma_\varepsilon^2 \neq 0$$ Hence, the covariance is different from zero if there is a linear regression relation between *X* and *Y*. The only way to void this problem is to force the variance of the measurement error to zero. This is of course difficult in practice.

The third case considers the combined effects of measurement errors in both the dependent and independent variables. This case add nothing new to the discussion since the effect will be the same as when just the explanatory variables contains measurement errors. That means that the OLS estimators are both biased and inconsistent, and the problem drives primarily from the error that comes from the explanatory variable.

8. Dummy variables

Until now all variables have been assumed to be quantitative in nature, which is to say that they have been continuous. However, many interesting variables are expressed in qualitative terms, such as gender, educational level, time periods and seasons, private or public and so forth. These qualitative measures have to be transformed into some proxy so that it could be represented and used in a regression. Dummy variables are discrete transformations and used for this purpose. They are artificial variables that work as proxies for qualitative variables and since they are discrete we need to be careful when working with then and how to interpret them. The purpose of this chapter is to describe different techniques on how to use dummy variables and categorical variables in general and how to interpret them.

Gender is a typical example of a qualitative variable that need to be transformed into a numerical form so that it could be used in a regression. Since gender could be male or female it is a categorical variable with two categories. We therefore need to decide what category the dummy should represent and what category that should be used as a reference. If the dummy variable should represent men, the discrete variable *D* would take the value 1 for men and the value 0 for all other values in the data set. It is therefore important to be sure that all other observations really represent what you want it to represent. A dummy variable for men could therefore be expressed in this way:

$$D = \begin{cases} 1 & \text{if a man} \\ 0 & \text{otherwise (if a woman)} \end{cases} \tag{8.1}$$

When running the regression you can treat the dummy variable *D* as any other variables included in the model. The variable *D* could take other numerical values than 1 and 0, for instance 9 and 7, and it will not have any effect on its coefficient. However, the interpretation is easiest when using 1 and 0, which is the reason why you should follow the structure of (8.1) which is standard.

8.1 Intercept dummy variables

The most basic form of application using dummy variables is when only the intercept is affected. Using the categorical variable defined by (8.1) we can form the following model with two explanatory variables.

$$Y = B_0 + B_1 D + B_2 X + U \tag{8.2}$$

As can be seen from (8.1) *D* takes only two values. If we form the conditional expectation with respect to the two categories of *D* we receive:

$$E[Y \mid D = 1, X] = B_0 + B_1 + B_2 X \tag{8.3}$$

$$E[Y \mid D = 0, X] = B_0 + B_2 X \tag{8.4}$$

The only thing that differs between the two expectations is the coefficient for the dummy variable. When *D*=1 we see that the conditional expectation in (8.3) consist of two constants B_0 and B_1 which sum represents the intercept in that case. However, when *D*=0, the conditional expectation will be given by (8.4) which only contain one constant B_0. Hence, the model as a whole contains two intercepts B_0 and B_0+B_1.

If we take the difference between the two conditional expectations we receive:

$$E[Y \mid D=1, X] - E[Y \mid D=0, X] = B_1 \tag{8.5}$$

which equals the coefficient for the dummy variable. Since our binary variable *D* is discrete, we can not take the derivative of *Y* with respect to *D*, since a derivative requires a continues variable, and therefore is undefined here. In order to find the corresponding marginal effect in this case we have to form the difference given by (8.5) and conclude that when *D* moves from 0 to 1, the conditional expectation of *Y* change by B_1 units, which represents the marginal effect for the linear model. When working with the linear model, it makes no difference if we treat the dummy variable as it was continuous when calculating the marginal effect since they become the same. But with other functional forms it makes a difference.

Example 8.1

Assume the following regression result from a model given by (8.2) with *Y* being the hourly wage rate, *D* a dummy for men, and *X* a variable for years of schooling. The dependent variable is expressed in Swedish kronor (SEK). Standard errors are given within parenthesis:

$$\begin{aligned} \hat{Y} = 55.9 &+ 21.9D + 2.4X \\ (8.16) &\quad (4.30) \quad (0.63) \end{aligned}$$

Use the regression results to calculate how much higher the average hourly wage rate is for men. First we have to check if the coefficient for the male dummy is significant. With a t-value equal to 5 the coefficient is significantly different from zero at any conventional significance levels. The marginal effect measured with this regression says that men earn 21.9 SEK/hours more than women do on average, controlling for years of schooling.

In the empirical human capital literature the functional form most often used is the log-linear, which means that our model would look like this:

$$\ln Y = B_0 + B_1 D + B_2 X + U \tag{8.6}$$

In order to find the marginal effect here, we have to remember that it is the effect on *Y* that is of interest, not ln*Y*. Therefore, the first step must be to transform the regression equation using the anti log and form the conditional expectation of *Y*. Doing that we receive:

$$\hat{Y} = e^{(B_0 + B_1 D + B_2 X + \sigma_U^2 / 2)} \tag{8.7}$$

where σ_U^2 represents the population variance of the error term. Hence, in order to receive the conditional expectation given by (8.7) we have to assume that *U* is a normally distributed variable, with mean zero, and variance equal σ_U^2. When that is the case, it is possible to show that $E\left[e^U\right] = e^{\sigma_U^2 / 2}$.

Had *D* been continuous, would B_1 have represented the relative change in *Y* from a unit change in *D*. Since it is not continuous, we have to form the relative change in *Y* using the conditional expectation given by (8.7) instead. Doing that we receive:

Marginal effect:

$$\frac{E[Y \mid D=1, X] - E[Y \mid D=0, X]}{E[Y \mid D=0, X]} = \frac{e^{\left(B_0+B_1+B_2X+\sigma_U^2/2\right)} - e^{\left(B_0+B_2X+\sigma_U^2/2\right)}}{e^{\left(B_0+B_2X+\sigma_U^2/2\right)}} = e^{B_1} - 1 \qquad (8.8)$$

Hence, in order to find the relative change in the conditional expectation of *Y*, we simply use the estimated value of B_1 and apply the formula given above. In order to find the corresponding standard error of the marginal effect we simply apply a linear approximation to the non linear expression. If we do that we end up with the following formula:

The variance of the marginal effect: $V\left(e^{b_1} - 1\right) = V\left(e^{b_1}\right) = \sigma_{b_1}^2 \times \left(e^{b_1}\right)^2$ (8.9)

Example 8.2
Assume the following regression results, from a model given by (8.6), with *Y* being the hourly wage rate, *D* a dummy for men, and *X* a variable for years of schooling. The dependent variable is expressed in Swedish kronor (SEK). Standard errors are given within parenthesis:

$$\begin{aligned}\ln \hat{Y} &= 4.02 + 0.18D + 0.03X \\ &\quad (0.03) \quad (0.02) \quad (0.01)\end{aligned}$$

Since we are interested in the marginal effect of *D* on *Y*, we have to calculate it using the regression results. By (8.8) and (8.9) we receive:

Marginal effect: $e^{b_1} - 1 = e^{0.18} - 1 = 0.197$

Standard error: $\sqrt{\left(e^{b_1}\right)^2 \times \sigma_{b_1}^2} = e^{b_1} \times \sigma_{b_1} = e^{0.18} \times 0.02 = 0.024$

The t-value for the marginal effect equals 8.2, and is well above the critical value of any conventional level of significance. This implies a positive and significant marginal effect of 19.7 percent. That is, men earns on average 19.7 percent more per hour than women, controlling for education.

Observe that the estimated value is very close to the calculated relative change given by (8.8). It turns out that when the estimated coefficient is lower than 0.3 in absolute terms, the coefficient it self is a very good approximation to the exact value given by (8.8), and is therefore often used directly as such.

Observe that

$$e^{b_1} - 1 \approx b_1 \quad \text{when } |b_1| < 0.3$$

therefore researcher often use b_1 directly instead of the calculated value given by (8.8).

8.2 Slope dummy variables

As could be seen in the previous section, the dummy variable could work as an intercept shifter. Sometimes it is reasonable to believe that the shift should take place in the slope coefficient instead of the intercept. If we go back to the human capital model it is possible to argue that the difference in wage rate between men and women could be due to differences in their return to education. This would mean that men and women have slope coefficients that are different in size.

A model that control for differences in the slope coefficient for different categories of the qualitative variable could be expressed in the following way:

$$\begin{aligned}\ln Y &= B_0 + (B_1 + B_2 D)X + U \\ &= B_0 + B_1 X + B_2 (DX) + U \qquad (8.10)\end{aligned}$$

In this case the slope coefficient for X equals B_1 when D=0 and B_1+B_2 when D=1. Hence, a way to test if the return to education differs between men and women would be to test if B_2 is different from zero, which should be tested before going on to test if B_1+B_2 is different from zero. Observe that the coefficient for the cross product is interpreted differently if both variables had been continuous. Since D is binary, DX is only active when D=1, and the corresponding effect is therefore related to the category specified by D=1.

Example 8.3
Use the same data set as in Example 8.2 and estimate the coefficients in (8.10). The results are presented below with standard errors within parenthesis:

$$\begin{aligned}\ln\hat{Y} &= 4.11 + 0.024X + 0.014DX \\ &\quad (0.031)(0.003) \quad (0.001)\end{aligned} \tag{8.11}$$

Use the regression results to investigate if there is a difference in the return to education between men and women. To answer that question we simply test the estimated coefficient for the cross product. Doing that, we receive a t-value of 10.3 which is above any critical values of conventional significance level. Hence, if this specification is correct, we can conclude that the returns to education differ between men and women.

8.2.1 A model will intercept and slope dummy variable

Whenever working with cross products it is very important to always include the involved variables separately to separate that kind of effect from the cross product. If it is the case that D in itself has a positive effect on the dependent variable, that unique effect will be part of the cross effect otherwise. Hence whenever including a cross product the model should be specified in the following way:

$$\ln Y = B_0 + B_1X + B_2D + B_3(DX) + U \tag{8.12}$$

When we include the two variables, X and D separately and together with their product we allow for changes in both the intercept and the slope. If it turns out that the coefficient of B_2 is not significant one can go on and reduce the specification to (8.10), but not otherwise.

Example 8.4
Extend the specification of (8.10) by including D separately. That is, estimate the parameters of the model given by (8.12) and interpret the results. Doing that, we received the following results, with standard errors within parenthesis.

$$\begin{aligned}\ln\hat{Y} &= 4.006 + 0.033X + 0.210D - 0.002DX \\ &\quad (0.045) \quad (0.004) \quad (0.062) \quad (0.005)\end{aligned} \tag{8.12}$$

By investigating the t-values we see that b_1 and b_2 are statistically significant from zero. But the t-value from the cross product is not significant any more. Since D alone has a significant effect on the dependent variable, there is little effect left from the cross product, and hence we conclude that there is no difference in the return to education between men and women.

Example 8.3 and 8.4 should convince you that it is very important to include the variables that appear in a cross product separately, since they might stand for the main effect. In Example 8.3 we did not include *D* even though it was relevant. In chapter 7 we learned that omitting relevant variables has consequences and bias the remaining coefficients. In this case it made us to draw the wrong conclusion about the return to education for men and women.

8.3 Qualitative variables with several categories

The human capital model described above includes a continuous variable for the number of years of schooling. When including a continuous variable for schooling it is under the belief that the hourly wages are set and determined based on this measure. An alternative approach would be to argue that it is the level of schooling, the received diploma, that matters in the determination of the wage rate. That calls for a qualitative variable with more than two categories. For instance:

$$D = \begin{cases} 0 & \text{Primary schooling} \\ 1 & \text{Secondary schooling} \\ 2 & \text{Post secondary schooling} \end{cases} \qquad (8.13)$$

In order to include *D* directly into a regression model we have to make sure that the effect of going from primary schooling to secondary schooling on the hourly wage rate is of the same size as going from secondary schooling to a post secondary schooling. If that is not the case we have to allow for differences in these two effects. There are at least two approaches to this problem.

The first and most basic approach is to create three binary variables; one for each educational level, in the following way:

$$D_1 = \begin{cases} 1 & \text{Primary schooling} \\ 0 & \text{otherwise} \end{cases} \quad D_2 = \begin{cases} 1 & \text{Secondary schooling} \\ 0 & \text{otherwise} \end{cases} \quad D_3 = \begin{cases} 1 & \text{Post secondary schooling} \\ 0 & \text{otherwise} \end{cases}$$

We can now treat D_1, D_2 and D_3 as three explanatory variables, and include them in the regression model. However, it is important to avoid the so called **dummy variable trap**. The dummy variable trap appears when the analyst tries to specify and estimate the following model:

$$\ln Y = B_0 + B_1 D_1 + B_2 D_2 + B_3 D_3 + B_4 X + U \tag{8.14}$$

It is a mathematical impossibility to estimate the parameters in (8.14) since there is no variation in the sum of the three dummy variables, since $D_1+D_2+D_3=1$ for all observations in the data set. Since the model only can contain one constant, in this case the intercept, we can not include all three dummy variables. The easiest way to solve this is to exclude one of them and treat the excluded category as a reference category. We re-specify the model in following way:

$$\ln Y = B_0 + B_2 D_2 + B_3 D_3 + B_4 X + U \tag{8.15}$$

That is, if D_1 is excluded, the other categories will have D_1 as reference. B_2 will therefore be interpreted as the wage effect of going from a primary schooling diploma to a secondary schooling diploma, and B_3 will represent the wage effect of going from a primary schooling diploma to a post secondary schooling diploma. In order to determine the relative effects you may use the transformation described by (8.8).

An alternative to exclude one of the categories is to exclude the constant term, which would give us a model that looks like this:

$$\ln Y = C_1 D_1 + C_2 D_2 + C_3 D_3 + B_4 X + U \tag{8.16}$$

The three dummy variables will then work as three intercepts in this model; one for each educational level. The coefficients can therefore not be interpreted as relative changes in this case.

Example 8.5
Estimate the parameters of (8.15) and (8.16) and compare and interpret the results.

Specification I: (8.15)
$$\begin{aligned} \ln \hat{Y} = &\ 3.929 + 0.154 D_2 + 0.295 D_3 + 0.009 X \\ &(0.043) \quad (0.024) \quad (0.022) \quad (0.001) \end{aligned} \tag{8.17}$$

Specification II: (8.16)
$$\begin{aligned} \ln \hat{Y} = &\ 3.929 D_1 + 4.083 D_2 + 4.224 D_3 + 0.009 X \\ &(0.043) \quad (0.034) \quad (0.036) \quad (0.001) \end{aligned} \tag{8.18}$$

The three dummy variables represent three educational levels, and X represents the age of the individual. The first thing to notice is that B_0=C_1, B_0+B_2=C_2 and B_0+B_2+B_3=C_3. Hence, the two specifications are very much related. Furthermore C_2-C_1=B_2 and C_3-C_1=B_3. With help from specification II, we can derive the effect of going from a high school diploma to a college diploma by taking the difference between C_3 and C_2 which turns out to be equal to 0.14, i.e. a 14 percent increase. However, that effect could also have been received by taking the difference between B_3 and B_2. For the obvious reason there should be no change in the effect of the other variables included in the model (B_4 in this example) when alternating between specification I and II.

8.4 Piecewise linear regression

Dummy variables are also useful when modeling a non linear relationship that can be approximated by several linear relationships, known as **piecewise linear relationships**. In Figure 8.1 we see an example of a piecewise liner relationship. A typical example of such a relationship would be related to the income tax, which often is progressive, that is, the more you earn the larger share of your income should be paid in tax.

Let say that we are interested in describing how the income tax paid (Y) is related to the gross household income (X) and we specify the following model:

$$Y = A + BX + U \tag{8.19}$$

In order to transform (8.19) into a piecewise linear regression we need to define two dummy variables that will describe on what linear section the household is located. We define:

$$D_1 = \begin{cases} 1 & \text{Gross income is in the interval } X_1 \le X \le X_2 \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{Gross income is greater than } X_2 \\ 0 & \text{otherwise} \end{cases}$$

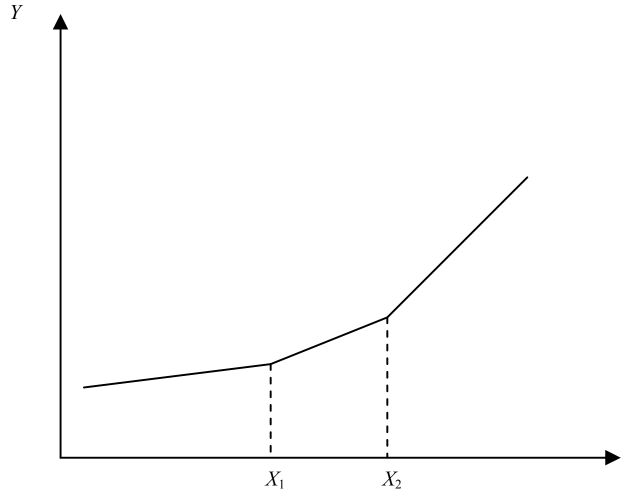


Figure 8.1 Piecewise linear regression

Next re-specify the intercept and the slope coefficient in (8.19) in the following way:

$$A = A_0 + A_1 D_1 + A_2 D_2 \tag{8.20}$$

$$B = B_0 + B_1 D_1 + B_2 D_2 \tag{8.21}$$

Substitute (8.20) and (8.21) into (8.19) and receive:

$$Y = (A_0 + A_1 D_1 + A_2 D_2) + (B_0 + B_1 D_1 + B_2 D_2)X + U$$

After multiply out the parenthesis we receive the following specification that could be used in estimation:

$$Y = A_0 + A_1 D_1 + A_2 D_2 + B_0 X + B_1 (D_1 X) + B_2 (D_2 X) + U \tag{8.22}$$

The estimated relations in the three income ranges are therefore given by:

When $X < X_1$	$\hat{Y} = a_0 + b_0 X$
When $X_1 \leq X \leq X_2$	$\hat{Y} = (a_0 + a_1) + (b_0 + b_1)X$
When $X > X_2$	$\hat{Y} = (a_0 + a_2) + (b_0 + b_2)X$

8.5 Test for structural differences

An important application using dummy variables is to test if the coefficients of the model differ for different sub groups of the population, or if they have changed over time. For instance, assume that we have the following wage equation expressed with a semi logarithmic (log-linear) functional form:

$$\ln Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{8.23}$$

with Y being the wage rate, X_1 the number of years of schooling, and X_2 the number of years of working experience. We would like to know if B_1 and B_2 differ between men (m) and women (w) simultaneously. That is, we would like test the following hypothesis:

$$\begin{aligned}&H_0 : B_{1m} = B_{1w}, B_{2m} = B_{2w}\\&H_1 : B_{1m} \neq B_{1w} \quad and/or \quad B_{2m} \neq B_{2w}\end{aligned}$$

In order to carry out this test using dummy variables we need to create an indicator variable D, for let's say for men, and then form the following regression model:

$$\ln Y = B_0 + B_1 D + B_2 X_1 + B_3 (X_1 D) + B_4 X_2 + B_5 (X_2 D) + U \tag{8.24}$$

Equation (8.24) will be representing the unrestricted model, where men and women are allowed to have different coefficients, and equation (8.23) will be representing the restricted case where men and women have the same coefficients. We will now compare the residual sums of squares (RSS) between the two models and use those in our test statistic given by:

$$F = \frac{(RSS_R - RSS_U)/df_1}{RSS_U / df_2} \sim F_{(J,n-k)} \tag{8.25}$$

If the RSS_R is very different from RSS_U we will reject the null hypothesis in favor of the alternative hypothesis. If they are similar in size, the test value will be very small and we say that the coefficients are the same for men and women.

Example 8.6

Assume that would like to know if the coefficients of equation (8.23) differ between men and women in the population. In order to test the joint hypothesis we need to run two regression models; one restricted model given by (8.23) and one unrestricted model given by (8.24). Using a sample of 1483 randomly selected individuals we received the following results:

Table 8.1 Regression results

	Restricted Model	Unrestricted Model
Residual Sum of Squares (*RSS*)	145.603	140.265
Degrees of freedom (*n*-*k*)	1483 – 3 = 1480	1483 – 6 = 1477

Using the results given in Table 8.1 we may calculate the test value using (8.25). The degrees of freedom for the numerator is calculated as the difference between degrees of freedom for the RSS from the restricted and

unrestricted model. That is 1480 – 1477 = 3. Another way to think about the degrees of freedom for the numerator is to express it in terms of the number of restrictions imposed by the restricted model compared to the unrestricted. The unrestricted model has 6 parameters, while the restricted model has only 3, which means that three parameters have been set to zero in the restricted model. Therefore we have 3 restrictions.
The test value using the test statistic is therefore equal to:

$$F = \frac{(RSS_R - RSS_U)/df_1}{RSS_U/df_2} = \frac{(145.603 - 140.265)/3}{140.265/1477} = \frac{1.7793}{0.095} = 18.73$$

The test value has to be compared with a critical value. Using a significance level of 5 percent the critical value equals 2.6. Hence, the test value is much larger than the critical value which means that we can reject the null hypothesis. We can therefore conclude that the coefficient of the regression model differ for men and women.

9. Heteroskedasticity and diagnostics

The classical assumption required for the OLS estimator to be efficient states that the variance of the error term has to be constant and the same for all observations. This is referred to as a homoskedastic error term. When that assumption is violated and the variance is different for different observations we refer to this as heteroskedasticity. This chapter will discuss the consequences of violating the homoskedasticity assumption, how to detect any deviations from the assumption and how to solve the problem when present.

9.1 Consequences of using OLS

The classical assumptions made on the error terms are that they are uncorrelated, with mean zero and constant variance σ_U^2. In technical terms this means that

$$E[U_i] = 0 \tag{9.1}$$

$$V[U_i] = \sigma_U^2 \tag{9.2}$$

$$Cov[U_i, U_j] = 0 \tag{9.3}$$

Assumptions (9.1) and (9.3) are in use to make the OLS estimators unbiased and consistent. Assumption (9.2) is important for the OLS estimator to be efficient. Hence, if (9.2) is ignored we can no longer claim that our estimator is the best estimator among linear unbiased estimators. This means that it is possible to find another linear unbiased estimator that is more efficient.

Heteroskedasticity implies that

- The OLS estimators of the population parameters are still unbiased and consistent.
- The usual standard errors of the estimated parameters are biased and inconsistent.

It is important to understand that the violation of (9.2) makes the standard errors of the OLS estimators and the covariances among them biased and inconsistent. Therefore tests of hypothesis are no longer valid, since the standard errors are wrong. To see this, consider the variance of the estimator for the slope coefficient of the simple regression model:

$$V(b_1) = V\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 V[Y_i]}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} \tag{9.4}$$

The expression given by (9.4) represents the correct variance that should be used. Unfortunately it involves the unknown population variance of the error term which is different for different observations.

Since the error term is heteroskedastic, each observation will have a different error variance. The expression will therefore deviate from the variance estimated under homoskedasticity, that is:

$$V(b_1) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sigma_i^2}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2} \neq \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = E\left[S_{b_1}^2\right] \tag{9.5}$$

As can be seen in (9.5) the variance of the OLS estimator is different from the expected value of the sample variance of the estimator that works under the assumption of a constant variance.

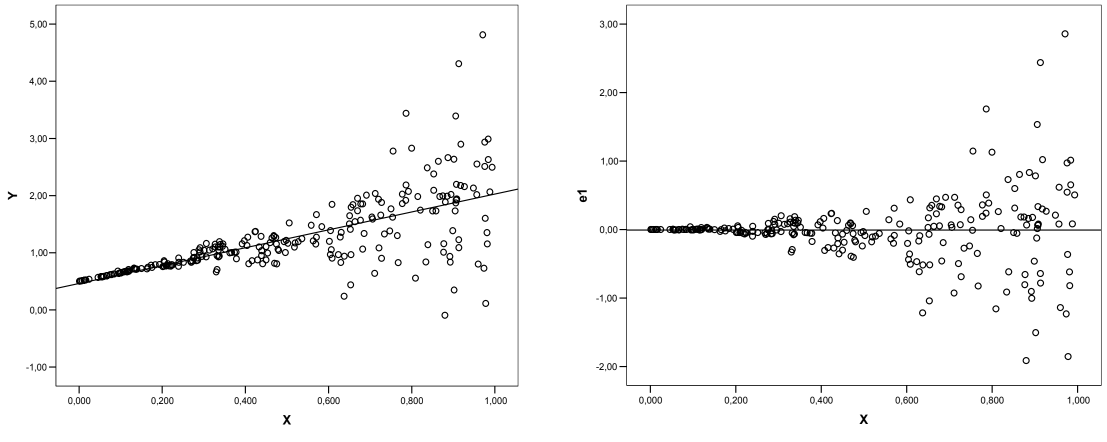
An important use of the regression equation is that of making predictions and forecasts of the future. Since the OLS estimators are unbiased and consistent, so will the forecasts. However, since the estimators are inefficient, the uncertainty of the forecasts will increase, and the confidence interval of the forecast will be biased and inconsistent.

9.2 Detecting heteroskedasticity

Since we know that heteroskedasticity invalidate test results it is very important to investigate whether our empirical model is homoskedastic. Fortunately there are a number of test, graphical as well as statistical that one can apply in order to receive an answer to the question. Below the most commonly used test will be discussed.

9.2.1 Graphical methods

A natural starting point in detecting possible deviations from homoskedasticity is to plot the data. Since we are interested in the behavior of the error term and its variation, two obvious scatter plots are given in Figure 9.1, and that comes from a simple linear regression model. In Figure 9.1a the dependent variable is plotted against its explanatory variable *X*. Here we can see a clear pattern of heteroskedastidity which is driven by the explanatory variable. That is, the larger the value of *X,* the larger is the variance of the error term.



a) Scatter plot of *Y* against *X* b) Scatter plot of *e* against *X*

Figure 9.1 Scatter plots to detect heteroskedasticity

As an alternative to Figure 9.1a the estimated residuals could also be plotted directly against X, as is done in Figure 9.1b. In the simple regression case with just one explanatory variable these two graph are always very similar. However, when using a multiple regression models the picture might be different, since the residual is a linear combination of all variables included in the model. Since it is the partial effect of X on the residual that is of primary interest, it is advised that the residual should be plotted against all involved variables separately. If it is possible to find a systematic pattern that give indications of differences of the variances over the observations, one should be concerned. Graphical methods are useful, but sometimes it is difficult to say if heteroskedasticy is present and found harmful. It is therefore necessary to use statistical test. The graphical method is therefore merely a first step in the analysis that can give a good picture of the nature of the heteroscedasticity we might have, which will be helpful later on when we are correcting for it.

Example 9.1

We are interested in the rate of the return to education and estimate the coefficients of the following human capital model:

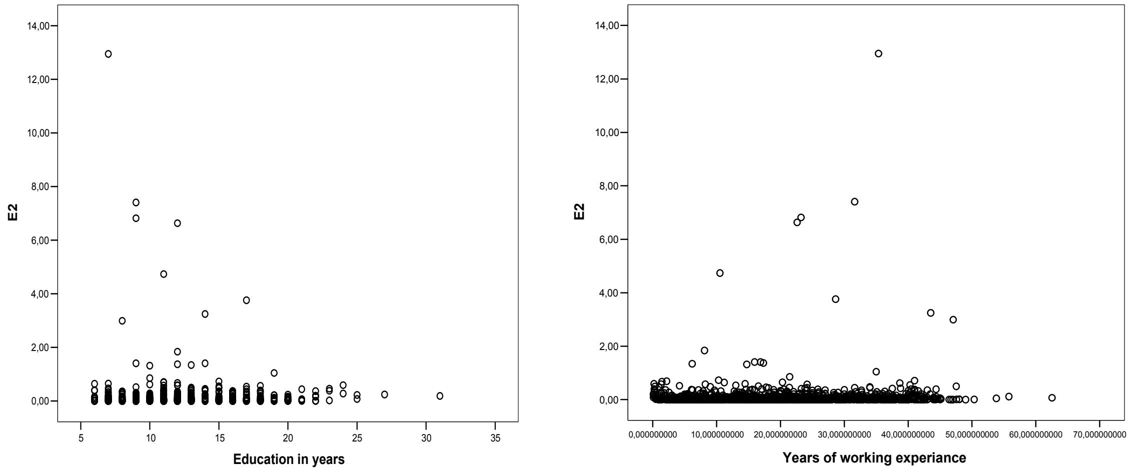
$$\ln Y = B_0 + B_1 ED + B2ED^2 + B_3 year + B_4 year^2 + U \tag{9.6}$$

We use a sample of 1483 individuals with information on hourly wages in logarithmic form (ln*Y*), years of schooling (*ED*), and years of work experience (*year*). Both explanatory variables are also squared to control for any non linear relation between the dependent variable and the two explanatory variables. Using OLS we received the following results with t-values within parenthesis:

$$\begin{array}{llllll} \ln\hat{Y} = 3.593 & +\,0.066ED & -\,0.001ED^2 & +\,0.017year & -\,0.000year^2 & \\ \quad (40.1) & \quad (5.2) & \quad (-2.2) & \quad (6.8) & \quad (-2.9) & \end{array} \qquad (9.7)$$

$\ln\hat{Y}$ should be interpreted as the predicted value of ln*Y*. We observe that all coefficients are significantly different from zero. The squared terms have very small coefficients, even though their t-values are sufficiently large to make them significant. Observe that the coefficient for the square of *year* is different from zero. Since the value is very small and we only report the first three decimals it appears to be zero. Its t-value shows that the standard error is even smaller.

We suspect that our residual might be heteroskedastic and we would like to investigate this by looking at a graph between the residual and the two explanatory variables. Sometimes, to enlarge the possible differences in variance among the residuals it is useful to square the estimated residual. If we do that we receive the graphs given in Figure 9.2.



a) Plot between e^2 and *ED* b) Plot between e^2 and *year*

Figure 9.2 Graphical analysis of error variance

In Figure 9.2a we see the squared error term against the number of years of schooling, and a pattern can be identified. The error variance seems to be larger for lower years of schooling than for more years of schooling. This is of course just an indication that we need to investigate further using formal statistical test.

In Figure 9.1b the picture is less obvious. If ignoring the top outlier that makes it look likes the variance is higher around 35 year of work experience, it is difficult to say if there is some heteroskedasticity to worry about. Since it is an unclear case, we still need to investigate the issue further, holding in mind that hypothesis testing is meaningless if the standard errors of the parameter estimates are wrong. Below we will go through the basic steps of the three most popular tests discussed in the textbook literature.

9.2.2 Statistical tests

The three most common statistical test procedures to identify a problem of heteroskedasticity are the Goldfeld-Quant test, the Breusch-Pagan test, and the White's test. Below we will shortly describe the logic of the tests and how they are implemented.

The Goldfeld-Quant test (GQ) works under the assumption that the error variance is equal for all observations, which is to say that the error term is homoskedastic. When this is true, the variance of one part of the sample must be the same as the variance of another part of the sample independent on how the sample is sorted. If this is not the case we must conclude that the data at hand is heteroskedastic. The following basic steps complete the GQ-test:

1) Sort the sample according to a variable that you believe drives the size of the variance. If the variable X_1 is related to the size of the variance, sort the data set in an increasing order of X_1, and divide the sample into three groups of equal size and omit the middle group. If the sample size is very small (i.e. each group is less than 100 observations), it is enough to divide the sample into two groups without omitting any observations.

2) Run the model for each sub sample and calculate the Residual Sum of Squares (RSS) for each group:

$$RSS_1 = \sum_{i=1}^{n_1} e_i^2 \qquad RSS_2 = \sum_{i=n_1+1}^{n} e_i^2 \tag{9.8}$$

3) Form the hypothesis that is to be tested:

$$\begin{aligned} H_0 &: \sigma_i^2 = \sigma^2 \\ H_1 &: \sigma_i^2 = \sigma^2 X_{1i} \end{aligned} \tag{9.9}$$

4) Use the two Residual Sum of Squares to calculate the variance of the two sub samples and form the test function:

$$F = \frac{S_1^2}{S_2^2} = \frac{RSS_1/(n_1-k)}{RSS_2/(n_2-k)} \sim F_{(n_1-k,n_2-k)} \tag{9.10}$$

As a rule of thumb, one should always put the larger variance in the numerator. Choose a significance level and find the critical value to compare the test value with. If the test value is larger than the critical value you choose to reject the null hypothesis.

Example 9.2

We are going to investigate model (9.6) used in Example 9.1 to see if we can identify any heteroskedasticity using the GQ-test. In the graphical analysis we found an indication of heteroskedastisity related to the number of years of schooling. Therefore, we sort the data set in an increasing order of years of schooling and delete the 33 percent in the middle. We use the two remaining sub-samples and estimate a regression for each sample. Using the results from these regressions we calculate the corresponding variance for each regression:

$$S_1^2 = \frac{RSS_1}{n_1-k} = \frac{59.9641}{490} = 0.1223756 \qquad S_2^2 = \frac{RSS_2}{n_2-k} = \frac{43.6254}{489} = 0.08921339 \tag{9.11}$$

Using the estimated variances for the two sub samples we can calculate the test value:

$$F = \frac{S_1^2}{S_2^2} = \frac{0.1223756}{0.0892133} = 1.3717 \qquad (9.12)$$

Choosing a significance level of 5 percent we found a critical value equal to 1.16. Most statistical tables do not offer information on critical values for the degrees of freedom we have in this example. To approximate the critical value by using the numbers valid for infinity in both numerator and denominator would not be meaningful. Therefore we have been using Excel to calculate the critical value valid for the degrees of freedom in our case.

Since our test value is larger than the critical value we conclude that our model suffers from heteroskedasticity and that year of schooling is at least partly responsible. A general problem with this test is that it tends to reject the null hypothesis very often. That is, it is very sensitive to very small differences, especially when the degrees of freedom are in level with those that we have in this example, since that produce very small critical values.

In a second step you should also test the second variable, years of work experience, that could be part of the problem as well. However, we will not go through that here, and leaves that to the reader.

The Breusch-Pagan test (BP) is also a popular test procedure presented in most econometric text books. The BP-test is slightly more general than the GQ-test, since it allows for more than one variable at the time to be tested. The starting point is a set of explanatory variables that we believe drives the size of the variance of the error term. We will call them X_1, X_2, …, X_h, and we claim that the following specification could be a plausible specification for our error variance:

$$E\left[U_i^2\right] = \sigma_i^2 = \sigma^2 f\left(A_0 + A_1X_1 + A_2X_2 + ... + A_hX_h\right) \tag{9.13}$$

The variables included in (9.13) could be just a sub set of the explanatory variables of the model or it could be all of them. In Example 9.1 we could not be conclusive about whether just one or if both of our variables were driving the size of the variance. In a case like that it is advisable to included both the variables in the specification of the variance given by (9.13). The functional form is not expressed explicitly in (9.13) as stated, but we are going to use a linear specification, just as for the model we use.

The hypothesis of this test is:
$$\begin{aligned} &H_0 : A_1 = A_2 = ... = A_h = 0 \\ &H_1 : A_j \neq 0 \text{ for at leat one } j,\ j = 1,2,...,h \end{aligned} \tag{9.14}$$

In order to test the hypothesis we have to go through the following basic steps:

1) Run the regression for the model you believe suffers from heteroskedasticity using OLS.

2) Save the residuals and square them (e_i^2). Use the squared residual and run the following auxiliary regression:

$$e^2 = A_0 + A_1X_1 + A_2X_2 + ... + A_hX_h + \varepsilon \tag{9.15}$$

Equation (9.15) is a representation of (9.13) with a linear specification.

3) Even though it looks like we could use the classical approach of using an F-test to test the joint hypothesis, it turns out not to be possible since the dependent variable is a construction based on another model. Instead the following test statistic could be used to test the null hypothesis:

$$LM = nR_e^2 \sim \chi_h^2 \tag{9.16}$$

where n is the number of observations used in the regression of (9.15) and R_e^2 is the coefficient of determination received from (9.15). It turns out that the product of those two terms is chi-squared distributed with h degrees of freedom, where h is the number of restrictions, which in this case corresponds to the number of variables included in (9.15). The test value should therefore be compared with a critical value received from the Chi-square table for a suitable level of significance.

Example 9.3
In this example we will use the same data set and the same model as in Example 9.2. But this time the test will involve both the variables included in the model. We choose not to include the squared terms, even thought they in principle could be included. Following the basic procedure of the BP-test we specify and estimate the variance function with standard errors given within parenthesis:

$$\begin{aligned}&\hat{e}^2 = 0.063 - 0.001ED + 0.002year\\&\quad\ (0.060)(0.004)\qquad (0.001)\\&R_e^2 = 0.0028 \quad n = 1483\end{aligned}$$

Using this information we are able to calculate the test value:

$$LM = nR_e^2 = 1483 \times 0.0028 = 4.1524$$

Choosing a significance at the 5 percent level, the Chi-square table with 2 degrees of freedom shows a critical value of 5.99. Hence, the test value is smaller than the critical value and we are unable to reject the null hypothesis. This means that we have received a conflicting result compared with the GQ-test result. Since the GQ-test is very sensitive to small differences, we believe that the result of this test is more useful. However, the BP-test is a test that requires large data sets to be valid, and is sensitive to any violation of the normality assumption. Since we have more than 1000 observations we believe that our sample is sufficiently large, but in order to be sure we will move on with yet another common test called the White's test.

White's test is very similar to the BP-test but does not assume any prior knowledge of the heteroskedasticity, but instead examines whether the error variance is affected by any of the regressors, their squares or cross products. Therefore, it is also a large sample test but it does not depend on any normality assumption. Hence, this third test is more robust than the other two test procedure described above, and is sometimes also called the White's General Heteroskedasticity test (WGH). The basic steps in the procedure are as follows for a model with two explanatory variables, where (9.17) represents the main model and (9.18) the variance function that contains all the variables of the main function and their squares and cross products:

$$Y = B_0 + B_1X_1 + B_2X_2 + U \tag{9.17}$$

$$\sigma^2 = A_0 + A_1X_1 + A_2X_2 + A_3X_1^2 + A_4X_2^2 + A_5X_1X_2 + U \tag{9.18}$$

1) Estimate the parameters of equation (9.17) and create and save the residual.
2) Square the residual and run the auxiliary regression model given by (9.18).
3) Using the results from the auxiliary regression you can calculate the test value using (9.16). If the test value is larger than the critical value chosen, you reject the null hypothesis of homoskedasticity.

Example 9.4

We repeat the test executed in Example 9.3 and apply the WGH-test instead. Observe that the only difference is in the specification of the variance function. Following the basic steps given above we received the following results with standard errors reported within parenthesis:

$$\begin{aligned}&\hat{e}^2 = 0.230 - 0.023ED - 0.001year + 0.001ED^2 + 0.000year^2 + 0.000ED \times year\\&\quad\ (0.196)(0.024)\qquad (0.007)\qquad (0.001)\qquad (0.000)\qquad (0.000)\\&R_e^2 = 0.0039 \quad n = 1483\end{aligned}$$

Observe that the coefficients and their standard errors are different from zero even though some of them appear to be zero since they are expressed with just three decimal points. Their t-values are definitely different from zero.

Using these results we can calculate the test value:

$$LM = nR_e^2 = 0.0039 \times 1483 = 5.78$$

The critical value from the Chi-square table, with 5 degrees of freedom and a significance level of 5 percent, equals 11.07, which is larger than the test value. Hence this test confirms the conclusions from the previous test and we are unable to reject the null hypothesis of homoscedasticity. That is, we have no statistical material that points in the direction of heteroskedasticity.

9.3 Remedial measures

In the previous discussion we concluded that our error term was homoskedastic, or that the trace of any heteroskedasticity was not to worry about. However, if we have followed the suggestion by the graphical inspection and the GQ-test we would have believed that the heteroskedasticity could have been driven by one of the explanatory variables, which is one example of how heteroskedasticty could look like. It could also be the case that our model contains two different sub groups with different variances, so that there are two different variances to deal with. There is of course a number of different ways heteroskedasticty could be expressed. Below we will look at some examples were we correct for heteroskedasticity under the assumption of a specific form of heteroskedasticity.

When the nature of the heteroskedasticity is known, one can use Generalized Least Squares (GLS) to estimate the unknown population parameters. Below we will look at three different cases on how to transform the model so that GLS could be applied.

To run a regression using GLS instead of OLS is in practical terms the same thing, but we call it GLS when we have transformed the variables in the model so that the error term become homoskedastic. Below we will go through three cases under the assumption of using the following population regression model:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{9.19}$$

Case 1: $\sigma_i^2 = \sigma^2 X_{1i}$

The first thing to note is that we still assume that the expected value of the error term equals zero, which means that the variance of the error term may be expressed as $E\left[U_i^2\right] = \sigma^2 X_{1i}$. The objective with the transformation of the variables is to make this expectation equal to σ^2 and nothing more. If we accomplish this with just some transformations of the involved variables we are home free. Let us see how we can do that in this particular case.

We know that the variance of the error term is $V\left[U_i\right] = \sigma^2 X_{1i}$. Hence, if we divided everything in the model with the square root of the variable that the variance is proportional to, we would end up with a homoskedastic error term. To see this:

$$\frac{Y_i}{\sqrt{X_{1i}}} = B_0 \frac{1}{\sqrt{X_{1i}}} + B_1 \frac{X_{1i}}{\sqrt{X_{1i}}} + B_2 \frac{X_{2i}}{\sqrt{X_{1i}}} + \frac{U_i}{\sqrt{X_{1i}}} \tag{9.20}$$

In practice this is carried out by just transforming Y, X_1, and X_2 and creating a new constant equal to $1 \big/ \sqrt{X_{1i}}$, instead of 1 that use to be there next to B$_0$. Hence when running this specification in a computer you have to ask the software to run the regression through the origin, since we now have a model specific constant that moves with X_1. All computer software has that option, and once you have found how to do that, you just regress $\frac{Y_i}{\sqrt{X_{1i}}}$ on $\frac{1}{\sqrt{X_{1i}}}, \frac{X_{1i}}{\sqrt{X_{1i}}}, \frac{X_{2i}}{\sqrt{X_{1i}}}, \frac{U_i}{\sqrt{X_{1i}}}$. When you transform the variables in this way, you automatically transform the error term, which now is divided by the square root of X_1. Once that is done, we have a homoskedastic error term. That is

$$V\left(\frac{U_i}{\sqrt{X_{1i}}}\right)=\left(\frac{1}{\sqrt{X_{1i}}}\right)^2 V(U_i)=\frac{1}{X_{1i}}\sigma^2 X_{1i}=\sigma^2 \tag{9.21}$$

Observe that nothing happens with the parameter estimates. The only thing that happens is that the error term is transformed into a constant which will correct the standard errors for the parameters.

Case 2: $\sigma_i^2=\sigma^2 X_{1i}^2$

This case is very similar to the previous case with the exception that the variable X_1 is squared, which means that the variance increases exponentially with X_1. The argumentation is similar to the one we had above, and the objective is to receive a constant error term. Hence instead of dividing by the square root of X_1 we simply divide by X_1 it self. If we do that we receive:

$$V\left(\frac{U_i}{X_{1i}}\right)=\left(\frac{1}{X_{1i}}\right)^2 V(U_i)=\frac{1}{X_{1i}^2}\sigma^2 X_{1i}^2=\sigma^2 \tag{9.22}$$

Case 3: Two different variances

In this case we have an error term that takes only two values. Hence, our sample could include two groups with intrinsic differences in their variance. If these two groups are known, we can sort the data set with respect to these groups. For the first n_1 observations, which contains the first group, the error term has the variance σ_1^2 and for the remaining n_2 observations, corresponding to the second group, the error term has the variance σ_2^2. In order to solve the heteroskedasticity problem here, we need to estimate the two variances, by splitting the sample in two parts and estimate the regression variance separately for the two groups. Once that is done we proceed and transform as follows:

Step 1: Split the data set into two parts and estimate the model separately for the two sets of data:

$$Y_i=B_0+B_1X_{1i}+B_2X_{2i}+U_i,\quad \text{var}(U_i)=\sigma_1^2 \qquad i=1,\ldots n_1$$
$$Y_i=B_0+B_1X_{1i}+B_2X_{2i}+U_i,\quad \text{var}(U_i)=\sigma_2^2 \qquad i=n_1+1,\ldots,n$$

Step 2: Transform each section of the data set with the relevant standard deviation, and run the regression on the full sample of n observations using the transformed variables:

$$\frac{Y_i}{\sigma_1}=B_0\frac{1}{\sigma_1}+B_1\frac{X_{1i}}{\sigma_1}+B_2\frac{X_{2i}}{\sigma_1}+\frac{U_i}{\sigma_1} \qquad i=1,\ldots n_1$$
$$\frac{Y_i}{\sigma_2}=B_0\frac{1}{\sigma_2}+B_1\frac{X_{1i}}{\sigma_2}+B_2\frac{X_{2i}}{\sigma_2}+\frac{U_i}{\sigma_2} \qquad i=n_1+1,\ldots,n$$

By scaling the error term for each group using their standard deviation, the new transformed error term will have a variance that equals 1 in both sub samples. When merging the samples the total variance for the full model using all observations together will then be constant and equal to 1. To see this

$V\left(\frac{U_i}{\sigma_1}\right)=\frac{1}{\sigma_1^2}V(U_i)=1$ for i=1, …n_1 and $V\left(\frac{U_i}{\sigma_2}\right)=\frac{1}{\sigma_2^2}V(U_i)=1$ for i=n_1+1, …,n

Example 9.5
Assume that we would like estimate the parameters of the following model

$$Y = B_0 + B_1X_1 + B_2X_2 + U$$

and we know that the nature of the error variance is proportional to X_1 in the following way:

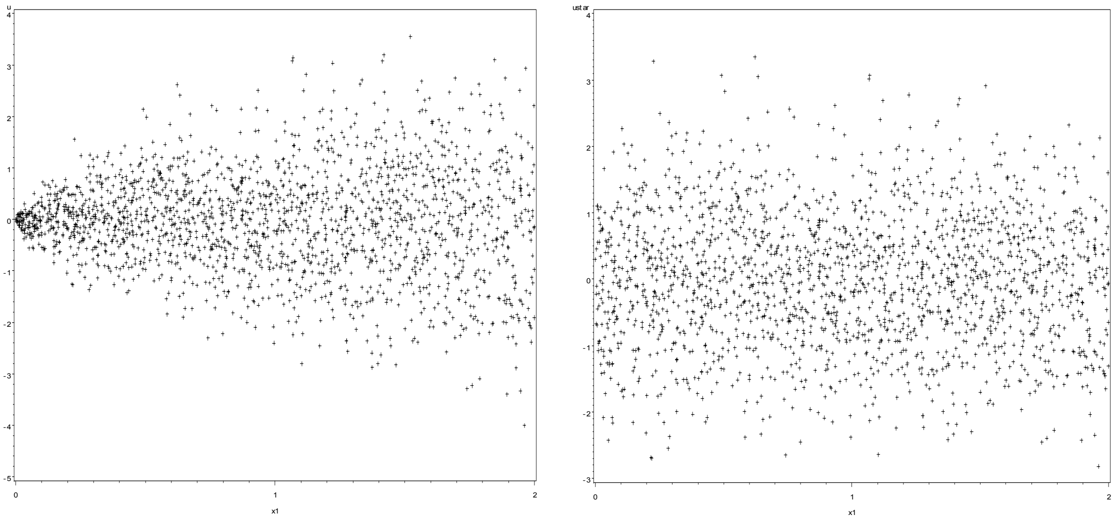
$$\sigma_i^2 = \sigma^2 X_{1i}$$

We would like to estimate the model using OLS and GLS and compare the results. Since we know how the structure of the heteroskedasticity, we apply GLS according to case 1.

Table 9.1 OLS and GLS estimates using 2000 observations

	Ordinary Least Squares		Generalized Least Squares	
	Coefficient	Standard error	Coefficient	Standard error
Constant	1.005	0.059	0.996	0.013
X_1	0.425	0.038	0.475	0.025
X_2	1.578	0.075	1.497	0.031
R^2	0.210		0.955	
MSE	0.956		0.991	

In Figure 9.3 we compare the residual plots before and after correcting for heteroskedasticity to see if the problem is fully solved. From Figure 9.3b the picture looks satisfying.



a) OLS residual plot versus X_1 b) GLS residual term versus X_2

Figure 9.3 Estimated residual plots before and after correction for heteroscedasticity

Table 9.1 show the results from OLS and GLS applied to a heteroskedastic model. As can be seen the estimated coefficient does not deviate that much which is what we expected since heteroskedasticity has no effect on the unbiasedness and consistency of the OLS estimator. When comparing the standard errors of the two estimations, large differences appear. The standard errors of the OLS estimated slope coefficients are twice as large as those for the corrected model. However the conclusions from the two estimations are the same. That is due to the relatively large sample that was used. If the sample would have been smaller, the corresponding t-values could have been much smaller and the chance of drawing the wrong conclusion would be greater. However, these results are sample specific. When the error term is heteroskedastic the standard errors are wrong and could be smaller or larger than the correct ones. So it is impossible to say something in advance with out knowing something about the exact nature of the heteroskedasticity.

Another important observation is related to the coefficient of determination. As can be seen it increased substantially. However, that does not mean that the fit of the model increased that much. Unfortunately it just means that after a transformation of the variables of the kind we did here, the coefficient of determination is of no use, since it is simply wrong.

9.3.1 Heteroskedasticity-robust standard errors

The approach of treating heteroskedasticity that has been described until now is what you usually find in basic text books in econometrics. But this approach is old fashion and researchers today tend to use a more convienient approach that is based on using an estimator for the standard errors that is robust to heteroskedasticity rather than doing all these investigations and then correct for it assuming a specific structure of the variance.

We know how the variance of the OLS estimator should look like for the simple linear regression model:

$$V(b_1) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sigma_i^2}{\left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right)^2} \tag{9.23}$$

Halbert White is an econometrician that showed that the unknown population variance could be replaced by the corresponding squared least square residual e_i^2 . By doing that one would receive consistent estimates of the true standard errors which provide a basis for inference in large samples. Hence, a heteroskedasticiy-consistent variance estimator could be estimated using the following formula:

$$S_{b_1}^2\Big|_{Robust} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 e_i^2}{\left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right)^2} \tag{9.24}$$

Since (9.24) is a large sample estimator it is only valid asymptotically, and test based on them are not exact and when using small samples the precision of the estimator may be poor. Fortunately there exist a small sample adjustment factor that could improve the precision considerably by multiplying the variance estimator given by *n*/(*n*-*k*). Furthermore and more importantly it is possible to generalize this formula to the multiple regression case, even thought it become slightly more complicated. Fortunately most econometric software such as STATA and SAS, includes the option of receiving robust standard errors together with the parameter estimates when running the regression. Hence in the practical work of your own you should always use the robust standard errors when running regression models.

Example 9.6
In this example we are going to use a random sample 1483 individuals and estimate the population parameters of the following regression function:

$$Y = B_0 + B_1 ED + B_2 ED^2 + B_3 Male + B_4 year + U$$

where *Y* represents the log hourly wages, *ED* the number of years of schooling, *Male* a dummy variable that indicates if the sample person is man, and *year* that represents the number of years of work experience. We are not sure whether we have a problem of heteroskedasticity and we therefore estimate the parameters with and without robust standard errors, to see how the estimates of the standard errors change. We received the following results:

Table 9.2 Regression results

Variables	OLS		Robust Estimation I		Robust Estimation II	
	P.E.	S.E.	P.E.	R.S.E.	P.E.	R.S.E.
Inctercept	3.646	0.087	3.646	0.105	3.815	0.041
Years of Education	0.063	0.012	0.063	0.016	0.037	0.003
Years of Education 2	-0.001	0.0004	-0.001	0.0006	-	-
Male (dummy)	0.123	0.017	0.123	0.017	0.124	0.0167
Years of Work exp.	0.008	0.001	0.008	0.001	0.008	0.001
RMSE	0.3079		0.3079		0.3083	

Note: P.E. stands for Parameter Estimates; S.E. stands for Standard Errors; R.S.E. stands for Robust Standard Errors. RMSE stands for Root Mean Square Error which is the standard deviation of the estimated residual.

Table 9.2 contains three regressions and the first column shows the results from the standard OLS regression assuming homoskedasticity. These results should be compared with the second column of estimates that use robust standard errors, which are heteroskedasticity consistent standard errors. Comparing those with the OLS case, we see that the robust standard errors are some what larger, which had consequences on the significance of the parameter for the squared education term, which no longer is significant. Including irrelevant variables in the regression makes the estimates less efficient. It therefore makes no sense to have the squared term included. In the third column, we re-estimate the model with out the squared term using robust standard errors.

Since we decided to use robust standard errors we could end up with a more parsimonious model, including only relevant terms. If we had included the squared education term, the marginal effects of education on earnings would be different and wrong. As can be seen from the RMSE measure that represents the estimated standard deviation of the error term it does not change very much among the specifications in Table 9.2. We should therefore conclude that the earnings model is not very sensitive to heteroskedasticity using this specification.

10. Autocorrelation and diagnostics

Autocorrelation or serial correlation often appears when working with time series data. One should understand that in order for autocorrelation to appear it is necessary that observations are correlated over a sequential order. In statistical terms this could be expressed as:

$$Cov\lfloor U_i, U_j \rfloor \neq 0 \quad \forall i \neq j \tag{10.1}$$

Hence, autocorrelation is a problem that frequently appears when working with data that has a time dimension. This means that it is meaningless to look for autocorrelation when working with cross sectional data which usually are based on random samples from a population, at a given point in time. This should be obvious since cross sectional data has no natural ordering that could generate a correlation. If correlation is found anyway, one can be sure that it is a fluke and has nothing to do with any underlying process.

As an example, we could think of a random sample of individuals taken from a population to analyze their earnings. To find a correlation between two randomly chosen individuals in this sample is not very likely. However if we follow the same individual over time, the correlation between par wise observations will be a fact, since it is the earnings of the same individual, and observed earnings for a given individual does not change very much between short time intervals.

This chapter will discuss the most important issues related to autocorrelation that an applied researcher need to be aware of, such as its effect on the estimated parameters when ignored, how to detect it and how to solve the problem when present.

10.1 Definition and the nature of autocorrelation

An autocorrelated error term can take a range of different specifications to manifest a correlation between pair wise observations. The most basic form of autocorrelation is referred to as the first order autocorrelation and is specified in the following way:

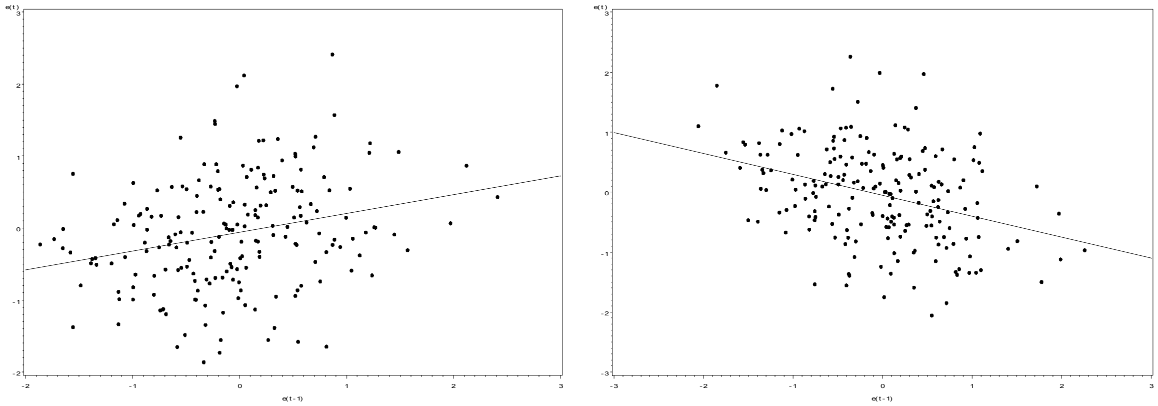
$$U_t = \rho U_{t-1} + V_t \qquad (10.2)$$

where *U* refer to the error term of the population regression function. As can be seen from (10.2) the error term at period *t* is a function of it self in the previous time period *t*-1 times the coefficient,ρ, which is referred to as the **first order autocorrelation coefficient** (This is the Greek letter rho, pronounced "row"). The last term *V*, is a so called white noise error term, and suppose to be completely random. It is often assume to be standard normal.

This type of autocorrelation is called autoregression because the error term is a function of its past values. Since *U* is a function of it self one period back only, as appose to several periods, we call it the first order autoregression error scheme, which is denoted AR(1). This specification can be generalized to capture up to *n* terms. We would then refer to it as the *n*th order of autocorrelation and it would be specified like this:

$$U_t = \rho U_{t-1} + \rho^2 U_{t-2} + ...\rho^n U_{t-1} + V_t \qquad (10.3)$$

The first order autocorrelation is maybe the most common type of autocorrelations and is for that reason the main target of our discussion. The autocorrelation can be positive or negative, and is related to the sign of the autocorrelation coefficient in (10.2). One way to find out whether the model suffer from autocorrelation and whether it is positive or negative is to plot the residual term against its own lagged value.



a) Positive autocorrelation, (+0.3) b) Negative autocorrelation (-0.3)

Figure 10.1 Scatter plots between e_t and e_{t-1}

Figure 10.1 present two plots that are two examples of how the plots could look like when the error term is autocorrelated. The graph to the left represents the case of a positive autocorrelation with a coefficient equal to 0.3. A regression line is also fitted to the dots in order to make it easier to see in what direction the correlation drives. Sometimes you are exposed to plots where the dependent variable or the residual term is followed over time. However, when the correlation is below 0.5 in absolute terms, it might be difficult to identify any pattern using those plots, and therefore the plots above are preferable.

10.2 Consequences

The consequences of having an autocorrelated error term are very similar to those that appear with a heteroskedastic error term. In short we have that:

1) The estimated slope coefficients are unbiased and consistent.

2) With positive autocorrelation the standard errors are biased and too small.

3) With negative autocorrelation the standard errors are biased and too large.

Since the expected value of the residual is zero, despite any autocorrelation, the estimated slope coefficients are still unbiased. That is, the property of unbiasedness and consistency does not require uncorrelated error terms. Confirm this by reading chapter 3 where we derived the sample estimators and discussed their properties.

The efficiency property of the OLS estimator does, however, depend on the assumption of no autocorrelation. To see this, it is useful to repeat how the variance of the slope estimator looks like in the simple regression case. Assume the following set up:

$$Y_t = B_0 + B_1 X_t + U_t \tag{10.4}$$

$$U_t = \rho U_{t-1} + V_t \tag{10.5}$$

$$V_t \sim N(0,1) \tag{10.6}$$

$$E(U_t) = 0 \text{ and } V(U_t) = \sigma^2 \tag{10.7}$$

With this setup we observe that the residual term, *U,* is autoregressive of order one. The covariance is therefore given by:

$$Cov(U_t, U_{t-1}) = Cov(\rho U_{t-1}, U_{t-1}) = \rho Cov(U_{t-1}, U_{t-1}) = \rho\sigma^2 \tag{10.8}$$

When generalizing this expression to an arbitrary distance between two error terms it is possible to show that it equals

$$Cov(U_t, U_{t-j}) = \rho^j \sigma^2 \tag{10.9}$$

With this set up, together with the knowledge from chapter 3 on how the variance of the OLS estimator looks like, we can examine the variance under the assumption of autocorrelation. The variance of the slope coefficient can be expressed in the following way

$$V(b_1) = \frac{1}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2} V\left(\sum_{i=1}^{n}(X_i - \overline{X})U_i\right)$$

$$= \frac{\sigma^2}{\left(\sum_{t=1}^{n}(X_t - \overline{X})^2\right)^2}\left(\sum_{t=1}^{n}(X_t - \overline{X})^2 + 2\rho\sum_{t=1}^{n}(X_t - \overline{X})(X_{t-1} - \overline{X}) + 2\rho^2\sum_{t=1}^{n}(X_t - \overline{X})(X_{t-2} - \overline{X}) + ...\right)$$

$$= \frac{\sigma^2}{\sum_{t=1}^{n}(X_t - \overline{X})^2}\left(1 + 2\rho\frac{\sum_{t=1}^{n}(X_t - \overline{X})(X_{t-1} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2} + 2\rho^2\frac{\sum_{t=1}^{n}(X_t - \overline{X})(X_{t-2} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2} + ...\right) \qquad (10.10)$$

If the autocorrelation coefficient were zero $(i.e.\ \rho = 0)$, the infinite series within the parenthesis in (10.10) would equal one. However, if ignoring the autocorrelation when present, we disregard this term which bias the variance of the slope coefficient. To receive a picture of how large in size the parenthesis is, it is useful to rewrite it into something more compact. In order to do that, we need to impose some assumptions on the behavior of *X*. We assume that the variance of *X* is constant and given by σ_X^2, and follows a first order autoregressive scheme, just like the error term of the model. This implies that $Cov(X_t, X_{t-j}) = r^j \sigma_X^2$, with *r* being the correlation coefficient for X_t and X_{t-1}. If we apply these assumptions to (10.10) we receive

$$V(b_1) = \frac{\sigma_U^2}{T\sigma_X^2}\left(1 + 2\rho r + 2\rho^2 r^2 + 2\rho^3 r^3 + ...\right) = \frac{\sigma_U^2}{T\sigma_X^2}\frac{1+\rho r}{1-\rho r} \quad \text{when } j \to \infty \tag{10.11}$$

In order to receive the compact expression given by (10.11) you have to know how to deal with geometric series. If you do not know that, do not worry! The important thing here is to see how the sign of the autocorrelation coefficient and the correlation between of X_t and X_{t-j} affect the size of the variance and induce a bias when ignoring autocorrelation. We know that both ρ and r takes values between -1 and 1, since they represents correlation coefficients. With this set up we can analyze the size of the adjustment factor due to autocorrelation. Let us investigate two basic and common cases:

1) $\rho > 0$ and $r > 0$ (Positive autocorrelation)

When this is true, the adjustment factor become: $\frac{1+\rho r}{1-\rho r} > 1$, which means that the usual estimates for the variance will be too small, and coefficients may appear more significant then they rely are. With a fixed value of *r*, the adjustment factor is increasing with the size of the autocorrelation coefficient, which increases the bias. If the value of *r* is zero, as it would be in cross sectional data for instance, the adjustment factor would be one, and the bias of the variance would be zero, independent of the size the autocorrelation coefficient. Most macro economic time series has an *r* value that is different from zero, and hence the case would in general not appear.

2) $\rho < 0$ and $r > 0$ (Negative autocorrelation)

With a negative autocorrelation, the adjustment factor become: $0 < \frac{1-\rho r}{1+\rho r} < 1$, which means that the usual estimates will be too large, and appear less significant then they rely are. With a fixed value of *r*, and an increasing value of the autocorrelation coefficient in absolute terms, the adjustment factor will be smaller, and increase the bias.

Hence, when we have autocorrelation amongst our residual terms, we get biased estimates of the standard errors of the coefficients. Furthermore, the coefficient of determination and the usual estimator for the error variance of the model will be bias as well. Autocorrelation is therefore a serious problem that needs to be addressed.

10.3 Detection of autocorrelation

From the previous discussion we understand that autocorrelation is bad which emphasize the importance of learning how to detecting it. Below we will describe the most common procedures found in the text book literature. We will not discuss any graphical methods since they sometimes are difficult to interpret. In the introduction of the chapter we gave some examples on how graphical methods could be used. In more advanced time series analysis, graphical methods based on autocorrelation functions and partial autocorrelation functions are used frequently. However, we will not discuss these methods here.

10.3.1 The Durbin Watson test

The Durbin Watson test (DW) is maybe the most common test for autocorrelation and is based on the assumption that the structure is of first order. Since first order autocorrelation is most likely to appear in time series data, the test is very relevant, and all statistical software has the option of calculating it automatically for you.

The Durbin-Watson test statistic for first order autocorrelation is given by:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T}(e_t)^2} \tag{10.12}$$

with *e* being the estimated residual from a sample regression model. To see that this test statistic is related to the first order autocorrelation case we may rewrite (10.12) in the following way:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T}(e_t)^2} = \frac{\sum_{t=2}^{T}e_t^2}{\sum_{t=1}^{T}(e_t)^2} + \frac{\sum_{t=2}^{T}e_{t-1}^2}{\sum_{t=1}^{T}(e_t)^2} - \frac{2\sum_{t=2}^{T}e_t e_{t-1}}{\sum_{t=1}^{T}(e_t)^2} \approx 1 + 1 + 2\hat{\rho} = 2(1-\hat{\rho}) \tag{10.13}$$

where $\hat{\rho}$ on the right hand side is the autocorrelation coefficient from an first order autoregression scheme. However, it is only an approximation since the expressions in the numerator sum from 2 to *T* instead of 1 to *T* as is the case in the denominator. The larger the value of *T* the better is the approximation.

From (10.13) it is possible to see that the *DW* test statistic only takes values between 0 and 4 since the autocorrelation coefficient only takes values between -1 and 1. Hence when the autocorrelation coefficient equals 0, the *DW* test statistics equals 2. If $DW > 2$ we have an indication of a negative autocorrelation, and if $DW < 2$ we would have an indication of a positive autocorrelation. However, since the relationship is an approximation, the *DW* test value can sometimes deviate from 2 even though the autocorrelation coefficient is zero. So the standard question is how much it is allowed to deviate? Could we use some critical values to help us interpret the estimated value of *DW*.

Unfortunately there exist no simple distribution function for this test function since it depends on the number of observations used as well as the values of the explanatory variables used in the regression. For that reason it is not possible to establish a precise critical value for the *DW* test statistic. However, Durbin and Watson made some simulations so that you, based on the number of observations used, and the number of parameters included in the model, can find a lower value (L) and an upper value (U) to compare the DW test value with.

Table 10.1 Possible outcomes from the Durbin-Watson test for autocorrelation

Positive AC	Inconclusive	No AC	Inconclusive	Negative AC
0 <*DW*< L	L <*DW*< U	U<*DW*<(4-U)	(4-U)<*DW*< (4-L)	(4-L) <*DW*< 4

Table 10.1 show five different regions where the DW-test value potentially could end up. If you receive a test value that is located in the interval between the lower value (L) and the upper value (U) your test is inconclusive and you have no use of the DW-test. However, if the DW-value is between 0 and the lower

value (L) you can draw the conclusion of having a positive autocorrelation. In the statistical table, with upper and lower values for the DW-test, you will only find the values that refer to the section below 2. In case of a negative autocorrelation you have to form the upper and lower value your self, using L and U as is done in Table 10.1.

Example 10.1
Assume that you have a time series with 150 observations, and two explanatory variables that will be used to explain the dependent variable. Running the regression you received a *DW*-test value equal to 1.63. In the table with critical values for the Durbin Watson test you found that L=1.71 and U=1.76. Since the test value is outside the inconclusive interval and below the lower value we have to draw the conclusion that our model suffer from positive autocorrelation.

Example 10.2
We have the same set up as in the previous example but with a *DW*-test value equal to 2.35. In the table we found the value for L=1.71 and the value for U=1.76. Using these values we can calculate the inconclusive interval related to *DW*-values larger than 2. Using the information in Table 10.1 we received: (4-U) = (4-1.76) = 2.24 and (4-L) = (4-1.71) = 2.29. Since the test value of 2.35 is outside the interval and larger than the upper value of 2.29 we must conclude that our model suffer from negative autocorrelation.

10.3.2 The Durbins h test statistic

As been described above, the *DW*-test is made for the purpose of testing for first order autocorrelation. Furthermore, it assumes that none of the explanatory variables are lagged dependent variables which would be the case when estimating a dynamic model. When that is the case the *DW*-test has a tendency to be close to 2 even though the error terms are serially correlated. Hence, the DW-test should not be used with the following kind of model:

$$Y_t = B_0 + B_1 Y_{t-1} + B_2 X_t + U_t \tag{10.14}$$

Fortunately there is an easy alternative to the *DW*-test that could be seen as a modified version of it and for that reason is called the Durbins *h* statistic. It is defined in the following way:

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{T}{1 - T\left[Var(b_1)\right]}} \tag{10.15}$$

where *DW* is the standard *DW*-test, *T* the number of observations and $Var(b_1)$ the square of the standard error of the estimated parameter for the lagged dependent variable. The test statistic has been shown to be standard normally distributed under the null hypothesis of no autocorrelation, which means that the test value should be compared with a critical value from the standard normal table.

The presence of autocorrelation in models that include lagged dependent variables is even more affected than the standard model. When the error term is serially correlated in a dynamic model the estimated parameters are biased and inconsistent. It is therefore very important to correct for the problem before using the estimates for anything.

Example 10.3
Assume that we have estimated the parameters of a dynamic model and received the following results with standard errors within parenthesis:

$$\begin{aligned}\hat{Y}_t &= \underset{(0.321)}{1.324} + \underset{(0.043)}{0.789}Y_{t-1} + \underset{(0.032)}{0.821}X_t \end{aligned} \qquad DW = 1.529$$

We use quarterly data over a 30 years period, which means that T=120. Since our model includes a lagged variable we lose one observation in the estimation. Using the information from the regression, we may form the Durbins *h* statistic:

$$h = \left(1 - \frac{1.529}{2}\right)\sqrt{\frac{119}{1 - 119 \times (0.043)^2}} = 2.925$$

Using a one sided test at the 5 percent significance level we receive a critical value of 1.645. Since the test value is much larger than the critical value, we must conclude that our error terms are serially correlated.

10.3.3 The LM-test

The LM test is a more general test of autocorrelation compared to the two previous tests. Furthermore, it allows for a test of autocorrelation of higher order than one, and can be used even though lagged dependent variables are included in the model. However, the LM test is a large sample test which means that it should be treated as an approximation when using small samples, compared to the DW-test that could be seen as an exact test.

The LM test is executed with the following steps:

1) Estimate the parameters of your main model: $Y_t = B_0 + B_1 X_t + U_t$ (10.16)

2) Create the residual term using the estimated parameters and lag it.

3) Extend your original model by including the estimated lagged residual in the specification:
$$Y_t = B_0 + B_1 X_t + \rho e_{t-1} + V_t \qquad (10.17)$$

4) Test the null hypothesis $H_0 : \rho = 0$ using a simple t-test. If you reject the null hypothesis you can conclude that you have autocorrelation.

The equation given by (10.17) can be extended to include more lags of the residual terms in order to test for higher order of autocorrelation.

Example 10.4
Assume that we have a time series model with two explanatory variables, and we suspect that the error term might be serially correlated of the second order.

$$Y_t = B_0 + B_1 X_{1t} + B_2 X_{2t} + U_t \qquad (10.18)$$

Since the error term is unobserved we need to estimate the residuals for the model using the estimated parameters of the model:

$$e_t = Y_t - b_0 - b_1 X_{1t} - b_2 X_{2t} \qquad (10.19)$$

Lag the estimated residuals from (10.19), re-specify equation (10.18) and receive:

$$Y_t = B_0 + B_1 X_{1t} + B_2 X_{2t} + \rho_1 e_{t-1} + \rho_2 e_{t-2} + V_t \qquad (10.20)$$

We estimated the parameters of the extended version of the model given by (10.20) and received the following estimates with standard errors within parenthesis:

$$\begin{aligned} \hat{Y}_t &= \underset{(1.301)}{3.241} + \underset{(0.091)}{0.378} X_{1t} + \underset{(0.192)}{0.562} X_{2t} + \underset{(0.029)}{0.324} e_{t-1} + \underset{(0.101)}{0.124} e_{t-2} \end{aligned} \qquad (10.21)$$

By investigating the significance of the coefficients of the two residual terms, we see that the first one is significantly different from zero, while the other is not. We therefore conclude that the residual term is serially correlated of the first order only.

Since we included lagged variables in the specification we lose some observations, and when including two estimated residuals as in (10.21) we loose two. If we have a large sample, losing two observations is not a big deal. However, if we only have 20 observations, loosing the first two observations might have an effect. One way to deal with this problem is to impose some values for e_0 and e_{-1}. Since the expected value of the residual terms equal zero, the missing observations could be replaced by zeros. Running the regression with and without the imposed values will give you an indication if the two missing observations are important. Other more advanced methods might be available.

10.4 Remedial measures

Once we have found that our error term is serially correlated we need to correct for it before we can make any statistical inference on the population. As for the case of heteroskedasticity, we need to transform the involved variables and therefore use generalized least square. The transformation looks different dependent on the order of autocorrelation. We will therefore look at the two most frequently used error structures, AR(1) and AR(2), and show how it should be done for those two cases. After that it should be an easy task to generalize the transformation method for an autoregression of order *n*.

10.4.1 GLS when AR(1)

The transformation will be explained by an example. Assume that the objective is to transform the following model:

$$Y_t = B_0 + B_1 X_t + U_t \tag{10.22}$$

For simplicity reasons we use the specification of the simple regression model. However, the method can be generalized to any number of explanatory variables. The objective is to transform the autocorrelated U_t with something that free from autocorrelation V_t. Assume that U_t is autoregressive of order one and given by:

$$U_t = \rho U_{t-1} + V_t \tag{10.23}$$

If we substitute (10.23) into (10.22) we receive:

$$Y_t = B_0 + B_1 X_t + \rho U_{t-1} + V_t \tag{10.24}$$

Form the following expression using (10.22):

$$\rho U_{t-1} = \rho Y_{t-1} - \rho B_0 - \rho B_1 X_{t-1} \tag{10.25}$$

Substitute (10.25) into (10.24) and rearrange:

$$\left(Y_t - \rho Y_{t-1}\right) = B_0(1-\rho) + B_1\left(X_t - \rho X_{t-1}\right) + V_t \tag{10.26}$$

Equation (10.26) is the transformed equation we are looking for. The error term of the original model is now replaced by V_t that is free from autocorrelation and we can estimate the regression equation using OLS. OLS, in combination with a variable transformation that results in a corrected error term, is what we call GLS.

10.4.2 GLS when AR(2)

The corresponding transformation in the AR(2) case is very similar. In this case our error term has the following shape:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + V_t \tag{10.27}$$

With that in mind we can extend equation (10.26) in the following way:

$$\left(Y_t - \rho_1 Y_{t-1} - \rho_2 Y_{t-2}\right) = B_0(1-\rho_1-\rho_2) + B_1\left(X_t - \rho_1 X_{t-1} - \rho_2 X_{t-2}\right) + V_t \tag{10.28}$$

The whole description above is based on the idea that the autocorrelation coefficient has been known. That is never the case and therefore it must be estimated. The estimated value is often received when you test for autocorrelation. In the Durbin Watson case the test statistic equal

$$DW = 2(1-\rho) \quad \Rightarrow \quad \rho = 1 - \frac{DW}{2} \tag{10.29}$$

This means that you can use the Durbin Watson test statistic to receive an estimated of the autocorrelation according to (10.29).

In case of higher order of autocorrelation the LM test should be applied. The coefficients in front of the lagged residual terms in (10.21) are estimates of the coefficients in (10.27). Those estimates could therefore be used when transforming the variables according to (10.28).

In the literature you will be able to find more advanced method to estimate the autocorrelation coefficient that could be used when applying GLS. However, statistical software, such as STATA and SPSS, will do most of the job for you. All you have to do is to specify the variables to be used in your model.

11. Multicollinearity and diagnostics

Multicollinearity refers to a situation with a high correlation among the explanatory variables within a multiple regression model. For the obvious reason it could never appear in the simple regression model, since it only has one explanatory variable. In chapter 8 we shortly described the consequences of including the full exhaustive set of dummy variables created from a categorical variable with several categories. We referred to that as to fall in the dummy variable trap. By including the full set of dummy variables, one end up with a perfect linear relation between the set of dummies and the constant term. When that happens we have what is called perfect multicollinearity. In this chapter we will in more detail discuss the issue of multicollinearity and focus on what sometimes is called imperfect multicollinearity which referrers to the case where a set of variables are highly correlated but not perfect.

Multicollinearity

> The lack of independence among the explanatory variables in a data set. It is a sample problem and a state of nature that results in relatively large standard errors for the estimated regression coefficients, but not biased estimates.

11.1 Consequences

The consequences of perfect correlation among the explanatory variables is easiest explained by an example. Assume that we would like to estimate the parameters of the following model:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{11.1}$$

where X_1 is assumed to be a linear combination of X_2 in the following way:

$$X_2 = a + bX_1 \tag{11.2}$$

and where a and b are two arbitrary constants. If we substitute (11.2) into (11.1) we receive:

$$\begin{aligned} Y &= B_0 + B_1 X_1 + B_2 (a + bX_1) + U \\ Y &= (B_0 + aB_2) + (B_1 + bB_2) X_1 + U \end{aligned} \tag{11.3}$$

Since (11.1) and (11.2) implies (11.3) we can only receive estimates of $(B_0 + aB_2)$ and $(B_1 + bB_2)$. But since these two expressions contain three unknown parameters there is no way we can receive estimates for all three parameters in (11.1). We simply need more information, which is not available. Hence, with perfect multicollinearity it is impossible to receive an estimate of the intercept and the slope coefficients.

This was an example of the extreme case of perfect multicollinearity, which is not very likely to happen in practice, other than when we end up in a dummy variable trap or a similar situation. More interesting is to investigate the consequences on the parameters and their standard errors when high correlation is present. We will start this discussion with the sample estimator of the slope coefficient B_1 in (11.1) under the assumption that X_1 and X_2 is highly correlated but not perfect. The situation for the sample estimator of B_2 is identical to that of B_1 so it is not necessary to look at both. The sample estimator for B_1 is given by:

$$b_1 = \frac{\left(r_{Y1} - r_{12}r_{Y2}\right)}{(1-r_{12}^2)}\frac{S_Y}{S_1} \tag{11.4}$$

The estimator b_1 is a function of r_{Y1} which is the correlation between Y and X_1, r_{12} the correlation between X_1 and X_2, r_{Y2} the correlation between Y and X_2, S_Y and S_1 which are the standard deviations for Y and X_1 respectively.

The first thing to observe is that r_{12} appears in both the numerator and the denominator, but that it is squared in the denominator and makes the denominator zero in case of perfect correlation. In case of a strong correlation, the denominator has an increasing effect on the size of the expression but since the correlation coefficient appears in the numerator as well with a negative sign, it is difficult to say how the size of the parameter will change, without any further assumptions. However, it can be shown that the OLS estimators remain unbiased and consistent, which means that estimated coefficients in repeated sampling still will center around the population coefficient. On the other hand, this property says nothing about how the estimator will behave in a specific sample. Therefore we will go through an example in order to shed some light on this issue.

Example11.1
Consider the following regression model:

$$Y = B_0 + B_1X_1 + U$$

We would like to know how the estimate of B_1 changes when we include another variable X_2 that is highly correlated with X_1. Using a random sample of 20 observations we calculate the following statistics.

$$S_Y = 5.1 \qquad r_{Y1} = 0.843$$
$$S_1 = 5.0 \qquad r_{Y2} = 0.878$$
$$r_{12} = 0.924$$

For the simple regression case we receive:

$$b_1 = r_{Y1}\frac{S_Y}{S_1} = 0.843 \times \frac{5.1}{5.0} = 0.86\,.$$

For the multiple regression case when including both X_1 and X_2 we receive:

$$b_1^* = \frac{0.843 - 0.924 \times 0.878}{1 - 0.924^2} \times \frac{5.1}{5.0} = 0.211\,.$$

Hence, when including an additional variable the estimated coefficient decreased in size as a result of the correlation between the two variables. Is it possible to find an example where the estimator is increasing in size in absolute terms? Well, consider the case where X_2 is even more correlated with X_1, lets say that r_{12}=0.99. That would generate a negative estimate and the small number in the denominator will make the estimate larger in absolute terms. It is also possible to make up an examples where the estimator moves in the other direction. Hence, the estimated slope coefficient could move in any direction as a result of multicollinearity.

In order to analyze how the variance of the parameter estimates change it is informative to look at the equation for the variance. The variance of (11.4) is given by the following expression

$$V(b_1) = \frac{\sum_{i=1}^{n} e_i^2}{n-3} \frac{1}{\left(1 - r_{12}^2\right)\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} \tag{11.5}$$

When the correlation between X_1 and X_2 equals zero, will the variance of the multiple regression coefficient coincide with the variance for the coefficient of the simple regression model. However, when the correlation equals 1 or -1 the variance given by (11.5) will be undefined just as the estimated slope coefficient. In sum, the greater the degree of the multicollinearity, the less precise will be the estimates of the parameters, which means that the estimated coefficients will vary a lot from sample to sample. But make no mistakes; collinearity does not destroy the nice property of minimum variance among linear unbiased estimator. It still has a minimum variance, but minimum variance does not mean that the variance will be small.

It seems like the level of both the estimated parameter and its standard error are affected by multicollinearity. But how will this affect the ratio between them; the t-value. It can be shown that the computed t-value in general will decrease since the standard error is affected more strongly compared to the coefficient. This will usually result in non significant parameter estimates.

Another problem with multicollinearity is that the estimates will be very sensitive to changes in specification. This is a consequence from the fact that there is very little unique variation left to explain the dependent variable since most of the variation is in common between the two explanatory variables. Hence, the parameter estimates are very unstable and sometimes it can even result in wrong signs for the regression coefficient, despite the fact that it is unbiased. A wrong sign is referred to a sign that is unexpected according to the underlying theoretical model, or the prior believes based on common sense. However, sometimes we are dealing with inferior goods which means that we have to be careful with what we call “wrong” sign. Unexpected signs usually require more analysis to understand where it comes from.

11.2 Measuring the degree of multicollinearity

Three measures of the degree of multicollinearity are often suggested in the literature: the use of a correlation matrix, the Variance Inflation Factor (VIF), and the tolerance measure. All statistical measures have their limitations, and therefore it is always useful to use several measures when investigation statistical properties of a data set.

Assume that we would like to estimate the parameters of the following model:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + U \tag{11.6}$$

We suspect that the variables are highly correlated and would like to investigate the matter. A natural starting point would be to look at a simple correlation matrix to investigate the pair wise correlations between the variables. That is easily done in any statistical software. Having access to a random sample of 20 observations we received the following results:

Table 11.1 A correlation matrix for the explanatory variables in (11.6)

	X_1	X_2	X_3
X_1	1	0.924	0.458
X_2		1	0.085
X_3			1

As can be seen from Table 11.1 some of the variables are highly correlated with each other, such as X_1 and X_2. X_1 is also correlated with X_3 but to a much lower degree and the correlation between X_2 and X_3 is basically zero. From the results of the table we can be sure that that B_1 and B_2 will be difficult to estimate with any good precision, since the correlation between X_1 and X_2 will inflate their standard errors.

To further analyze the multicollinearity we turn to the next measure which is called the **Variance Inflation Factor (VIF).** It is defined in the following way:

$$VIF(b_i) = \frac{1}{1 - R_i^2} \tag{11.7}$$

where R_i^2 is the squared multiple-correlation coefficient. The squared multiple-correlation coefficient for a specific parameter is a measure of the linear strength between a variable X_i and the rest of the variables included in the model. The squared multiple-correlation coefficient is nothing else than the coefficient of determination received from a auxiliary regression made for each variable against the other variables in the model. That is

R_1^2 is received from: $X_1 = C_{10} + C_{11}X_2 + C_{12}X_3 + U$, $VIF(b_1) = \frac{1}{1 - R_1^2}$

R_2^2 is received from: $X_2 = C_{20} + C_{21}X_1 + C_{22}X_3 + U$, $VIF(b_2) = \frac{1}{1 - R_2^2}$

R_3^2 is received from: $X_3 = C_{30} + C_{31}X_1 + C_{32}X_2 + U$, $VIF(b_3) = \frac{1}{1 - R_3^2}$

When the model contains only two explanatory variables the squared multiple correlation will coincide with the squared bivariate correlation coefficient between the two variables in the model. If you look at (11.5) you will see that the variance inflation factor is included in that expression, and is the factor that is multiplied with the variance of the coefficient of the simple regression model. Hence it is a measure that relates to the case of no correlation, and how the variance is inflated by imposing the correlation. The expression for the variance in the case of more than two explanatory variables has a similar expression but with the squared multiple correlation coefficient instead.

VIF takes values from 1 up to any large number. The closer the multiple-correlation coefficient is to one, the larger the value of VIF. Part of the definition of VIF, is the other multicollinearity statistic called the tolerance. The tolerance measure is the denominator of the VIF expression. Since the square of the multiple-correlation coefficient is a coefficient of determination we could interpret it as such. That means that we have a measure of how large share of the variation in one variable that is explained by a group of other variables. Hence, 1 minus this share would be interpreted as how much of the variation that is left unique for the specific variable and could be used to explain the dependent variable.

When are VIF and the tolerance an indication of a multicollinearity problem? We can shed some lights on that question by an example. Let us go back to model (11.6) and check the VIF and tolerance condition for the variables in that case. Most statistical software has routines for this, and hence you should not need to run the auxiliary regression by your self. Using one such routine in SPSS we received the following regression results and collinearity statistics:

Table 11.2 Regression results for (11.6)

	P.E.	S.E.	VIF	Tolerance
Constant	117.085	99.782	-	-
X_1	4.334	3.016	708.843	0.00141
X_2	-2.857	2.582	564.343	0.00177
X_3	-2.186	1.595	104.606	0.00956
R^2	0.801			
Test of over all significance: F =21.516				
P-value	<0.0001			

From Table 11.2 we see that none of the coefficients are significantly different from zero. We also observe that the coefficient of determination is above 80 percent. That implies that 80 percent of the variation in the dependent variable is explained by the three explanatory variables included in the model. Furthermore, the test of overall significance of the model is highly significant which is in line with the measure of fit. The picture described here is a good example of the consequences of high correlation between the involved variables. It will blow up the standard errors of the model even though the model as such has explanatory power.

When you are exposed to a situation like this, you must go on and calculate some multicollinearity statistics such as VIF and the tolerance for each variable. In the table we see that VIF take large values for all variables if we compare to the case of no correlation that results in VIF=1. From the analysis of the pair wise correlations we know that the reason for X_3 to have a relatively large VIF number is mainly due to the correlation with X_2, since the correlation with X_1 is very weak. Even so, VIF for X_3 is 104 and the corresponding tolerance is as low as 0.009 which means that X_1 and X_2 only leave 0.9 percent unique variation left of X_3s total variation that can be used in explaining the variation in Y. The remaining variation was in common with the other two variables and hence must be disregared. One should observe that even though the par wise correlations are relatively low, their multiple-regression correlation is much higher, which emphasize the shortcomings of only looking at pair wise correlations.

11.3 Remedial measures

With a given specification and data set there is not much one can do about the multicollinearity problem. It could therefore be seen as a state of nature in which data offers no information about some hypothesis that could have been tested using t-tests for the parameters of the model.

Doing nothing is most often not a very attractive alternative. If the alternative of receiving more data is possible, it would be a good solution. It would not solve the multicollinearity problem, but the small unique variation that exist, will be based on more data and if the increase in the number of observations is large enough, it could help increasing the precision of the estimators. To receive more data, and sometimes very much more data, is often very costly and/or time consuming and therefore often not an alternative.

Another alternative would be to change the variable specification. One way of doing that would be to drop one of the variables. If we, in the first place, had an economic relevant specification we know that the estimated parameters will be biased and inconsistent if dropping a relevant variable. Hence, we would only replace on problem with another and this alternative is therefore in general not very attractive.

An alternative approach would be to rethink the model so that it could be expressed in an alternative way. One way of doing that could be to categorize one of the problematic variables. In our example discussed above it was problematic to include X_1 and X_2 in the same regression. But if we replace X_1, which is a continuous variable, with level indicators (dummy variables) instead it would hedge the strong correlation with X_2 and increase the precision of the estimates. However, that would mean a slightly different model, and we have to be willing to accept that.

In the literature there are other more or less restrictive methods described to handle this problem, and non of them very convincing in there way of reducing the problem. We will therefore not go into any of those more advanced techniques here, since they would require more statistical knowledge that is beyond the scope of this text. But remember, multicollinearity is a state of nature, and is therefore not something that you solve, but instead something that you have to live with.

12. Simultaneous equation models

One important assumption of the basic linear regression model is that the error term has to be uncorrelated with the explanatory variables. If the explanatory variables are in fact correlated with the error term, it would lead to inconsistent estimates of the parameters of the model. In this chapter we will relax this assumption by including additional equations to the model that explains where the correlation is coming from, and discuss the conditions that need to be fulfilled to receive consistent estimate.

12.1 Introduction

This chapter will only scratch the surface of the issues involved in estimating simultaneously equations and should therefore be seen as an introduction to the subject. In order for the explanatory variables to be correlated with the error term, they need to be considered random, which was not the case in the previous chapters. The assumption of random explanatory variables does not change anything related to the property of the OLS estimators but it allows for the possibility of being correlated with the error term.

What are the consequences of having the explanatory variable being correlated with the error term? The answer to that question is very similar to the case when we have measurement errors in the explanatory variables, which make the estimates bias and inconsistent. To see this, consider the following simple macro economic model of income determination:

$$Y_t = C_t + I_t \tag{12.1}$$

$$C_t = B_0 + B_1 Y_t + U_t \tag{12.2}$$

with Y_t being the national income, I_t investments, C_t the consumption expenditure and U_t a stochastic term. Equation (12.1) is an identity and an **equilibrium condition**. Hence this model is formulated under the condition of being in equilibrium, and the equation show how national income is related to consumption and investment in equilibrium. The second, equation given by (12.2), is a **behavioral equation** since it defines the behavior of the consumption expenditure in this economy. Equations with stochastic error terms are to be considered behavioral. Since Y_t and C_t are left hand variables in this system of equations, their values are determined by the model. We therefore say that Y_t and C_t are **endogenous variables**. We have an additional variable included in the model which is the investment. Since it is a right hand variable in the consumption function we say that it is an **exogenous variable**, which is to say that the value of investments are determined outside the model, it is pre determined. Since investment is determined outside the model it is also uncorrelated with the stochastic term U_t.

The system of equations can be solved with respect to the two endogenous variables in order to receive their long run expressions. To solve the system with respect to Y_t, we substitute (12.2) into (12.1) and solve for Y_t. That results in the following expression:

$$Y_t = \frac{B_0}{1-B_1} + \frac{1}{1-B_1} I_t + \frac{1}{1-B_1} U_t \tag{12.3}$$

In order to receive the long run expression for consumption expenditure, we substitute (12.1) into (12.2) and solve for C_t:

$$C_t = \frac{B_0}{1-B_1} + \frac{B_1}{1-B_1} I_t + \frac{1}{1-B_1} U_t \tag{12.4}$$

With this setup it is easy to describe the consequences of estimating (12.2) ignoring the fact that it is part of a system. From (12.3) we can see that Y_t is a function of U_t which means it is correlated with U_t. Since Y_t is correlated with U_t, we can not use OLS to estimate the coefficients of (12.2) without bias. If consumption expenditure had not been part of this system one could have argued that Y_t and U_t in fact are uncorrelated. But when that is not the case we see from (12.3) how they are related.

It should know be obvious that the OLS estimators are biased in small samples due to the correlation between Y_t and U_t. But are they also inconsistent? That is, if we increase the number of observations to a very large number, will the estimators still be biased? To see this consider the OLS estimator for B_1:

$$b_1 = \frac{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right) C_t}{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right)^2} = B_1 + \frac{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right) U_t}{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right)^2} \tag{12.5}$$

This expression was developed in chapter 3 (see (3.12)). If we take the expected value of the estimator we will receive:

$$E[b_1] = B_1 + E\left[\frac{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)U_t}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2}\right] \tag{12.6}$$

The problem with the expectation on the right hand side is that Y_t is a random variable and correlated with U_t, and for that reason we can not proceed as in chapter 3. Furthermore, since the expectation is a linear operator we have that $E[A/B] \neq E[A]/E[B]$, which further complicates the problem. Even though this makes it clear that the estimator no longer is unbiased, we do not know how the second component on the right hand side of (12.6) behave in large samples. It can be shown that the limit of the OLS estimator is given by the following expression:

$$\lim_{t\to\infty} b_1 = B_1 + \frac{(1-B_1)\sigma_U^2}{\sigma_I^2 + \sigma_U^2} \tag{12.7}$$

(12.7) show that in the limit the sample estimator still deviate from the population parameter, which means that the bias remains in large samples.

Correlation between the error term and the explanatory variables in a single equation model using OLS would lead to:

- Biased and inconsistent parameter estimates
- Invalid tests of hypothesis
- Biased and inconsistent forecasts

12.2 The structural and reduced form equation

From the previous discussion we learned that when an equation belongs to a system of equation, estimating them separately using OLS would lead to biased and inconsistent estimates. Hence, in order to be able to estimate the parameters of the equation system it is important to consider the whole system since they interact with each other. Before going into issues of estimation we need to define some more concepts. Consider the following system of equations:

$$C_t = A_0 + A_1 Y_t + U_{1t} \tag{12.8}$$

$$I_t = B_0 + B_1 R_t + U_{2t} \tag{12.9}$$

$$Y_t = C_t + I_t + G_t \tag{12.10}$$

It is a macro economic model that extends the example from the previous section and is based on three equations. It is an income determination model, with two behavioral equations; one for consumption expenditure C_t, and one for net-investments I_t. The consumption function is a function of income Y_t and the investment function is a function of interest rate R_t. The income equation that specifies the equilibrium condition is a function of consumption, investment and government spending G_t. This model has three endogenous variables, C_t, I_t, and Y_t, and two exogenous variables R_t and G_t that are pre determined.

The system of equations given by (12.8)-(12.10) describes the structure of the economy that we would like to investigate. For that reason these equations are called **structural equations**. The coefficients of the structural equations represent the direct effect of a change in one of the explanatory variables. If we take (12.9) as an example, B_1 represents the marginal propensity to invest as a result from a change in the interest rate. This represents the **direct effect** of a change in interest rate on the net-investment.

Assume that we increase the interest rate. That will have a direct effect on the investments in this model, which in a second step via the equilibrium condition will have an effect on the income. The income in its term will affect the consumption level, and since income is endogenous it will have an effect the error term U_1 since they are correlated. The initial change in the interest rate, will in this way, affect the components in the system until the effect reaches its equilibrium level.

We can therefore talk about two types of effect; the short run effect and the long run effect. The long run effect can be received from the long run relationship that can be determined by solving the structural equation system with respect to the endogenous variables. To solve the system for Y_t we simply substitute (12.8) and (12.9) into (12.10) and solve for Y_t. If we do that we receive:

$$Y_t = \frac{A_0 + B_0}{1 - A_1} + \frac{B_1}{1 - A_1} R_t + \frac{1}{1 - A_1} G_t + \frac{U_{1t} + U_{2t}}{1 - A_1} \tag{12.11}$$

If we do the similar thing with respect to the other two endogenous variables we would receive the following expressions:

$$C_t = \frac{A_0 + A_1 B_0}{1 - A_1} + \frac{A_1 B_1}{1 - A_1} R_t + \frac{A_1}{1 - A_1} G_t + \frac{A_1 U_{2t} + U_{1t}}{1 - A_1} \tag{12.12}$$

$$I_t = B_0 + B_1 R_t + U_{2t} \tag{12.13}$$

By solving the structural system of equations with respect to the endogenous variables we have determined the **reduced form equations** for income, consumption and investment. The coefficients of the reduced form equations represent the full effect when the system is in equilibrium. The full effect of a change in interest rate on income is represented by $B_1/(1\text{-}A_1)$. It is also called the **interest rate multiplier** on income. There is a corresponding multiplier related to consumption and investments that can be found in their reduced form equations. Observe that the reduced form equation for investments only is a function of interest rate. Government spending does not have any affect on the investment, even though it has an effect on consumption and income.

The nice thing with the reduced form equations is that they may be estimated separately using OLS. That is, the coefficients in the reduced form equations can be consistently estimated using OLS. Since the structural parameters are part of the reduced form coefficients it is sometimes possible to indirect find the structural coefficient using the estimated values or the reduced form coefficients. For that to be possible, certain requirements need to be fulfilled. The structural coefficients must be exactly identified.

12.3 Identification

In order to be able to estimate the structural equation coefficients they need to be identified. So, what do we mean by that? To give an intuitive feeling for its meaning we will give an example before going into any formal and mechanical tests.

Consider the following two equation system:

$$Q = A_0 + A_1 P + A_2 X_1 + U_1 \qquad \text{(Supply)} \qquad (12.14)$$

$$Q = B_0 + B_1 P + U_2 \qquad \text{(Demand)} \qquad (12.15)$$

This system contains two endogenous variables (*P* and *Q*), and one exogenous (X_1) variable. These two equations represent a demand and supply system for a given market. The question is if any of these two equations are identified. That is to ask if the parameters of the two equations can be estimated consistently.

It turns out that the demand function is identified while the supply function is not. To see this, consider Figure 12.1. In order to identify the demand function we need some exogenous variation that could help us trace out the function. That could be done using the supply function. The supply function contains an exogenous variable X_1 and the supply function takes a new position for each value of X_1. In that process we identify the demand function. But in the demand function we have nothing unique that does not appear in the supply function so it is impossible to move the demand function while holding the supply function fixed. Hence it is the presence of an exogenous variable in one equation that allows us to estimate the parameters of the other equation. If X_1 had been included in both equations there would have been no unique variation in any of the equations and hence no equation had been identified. However, if another exogenous variable, X_2, had been introduced and placed in the demand function, we would receive some exogenous variation that could help us to identify the supply function. In that case both equations would have been identified.

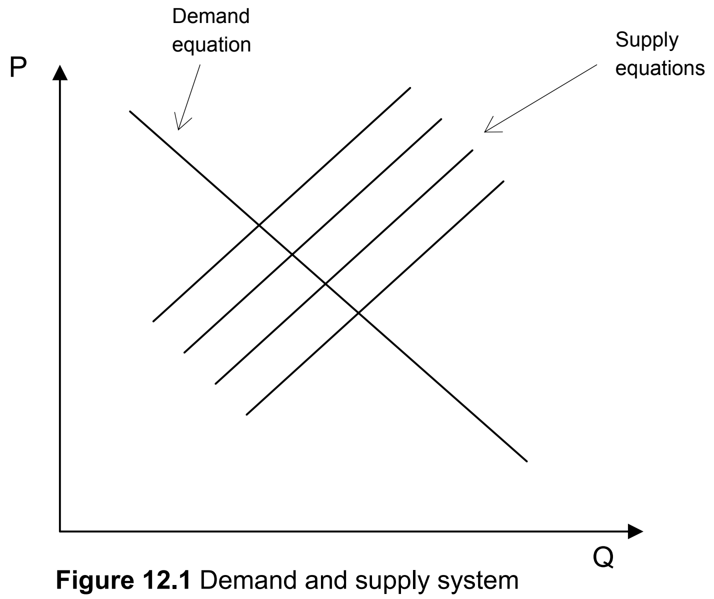


Figure 12.1 Demand and supply system

The process of identifying equations can be formalized in a decision rule that specify the conditions that have to be fulfilled in order to identify one or several equations in a system. In the literature two rules are described and one is slightly easier to use than the other.

12.3.1 The order condition of identification

The first decision rule for identification is the so called order condition. This rule specifies the necessary conditions for identification and is the more popular one of the two rules that will be discussed. Unfortunately it is not a sufficient rule, which means that it is possible that the equation is undefined even though the order condition says it is identified. However, in a system with only two equations, the order condition will work well and can be trusted.

Define the following variables:

M = The number of endogenous variables in the model

K = The number of variables (endogenous and exogenous) in the model excluded from the equation under consideration.

The order condition states that:

1) If $K = M - 1$ => The equation is exactly identified
2) If $K > M - 1$ => The equation is over identified
3) If $K < M - 1$ => The equation is under identified

When checking the order condition you have to do it for each equation in the system.

Example 12.1
Consider the following system:

$$Y_1 = A_0 + A_1 Y_2 + A_2 X_1 + U_1 \tag{12.16}$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + U_2 \tag{12.17}$$

Use the order condition to check if the equations are identified. In order to do that, we need to determine the value of M and K. This system contains two endogenous variables and the total number of variables, endogenous as well as exogenous, is 4.

For the first equation we have M-1=1 and K=1 since X_2 is excluded from (12.16). Since M-1=K we have that the first equation is exactly identified.

For the second equation we have M-1=1 and K=1 since X_1 is excluded from (12.17). Since M-1=K we have that also the second equation is exactly identified. When all the equations of the model are identified we say that the model is identified since we are able to estimate all the structural parameters.

Example 12.2
Consider the following system:

$$Y_1 = A_0 + A_1 Y_2 + A_2 X_1 + A_3 X_3 + U_1 \tag{12.18}$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + U_2 \tag{12.19}$$

In this example we have two endogenous variables and three exogenous variables with a total of five variables. M-1 will in this example equal 1 as before since we still have only two endogenous variables. Will the equations be identified in this case? The first equation contains four variables which means that one variable has been excluded from the equation, that is, X_2 does not appear in equation 1 and K=1. Since M-1=K the equation is exactly identified.

The second equation includes three variables which mean that two variables have been excluded. That is, X_1 and X_3 are not included in equation 2. That means that K=2, which means that M-1<K which leads to the conclusion that equation 2 is over identified.

12.3.2 The rank condition of identification

The rank condition is slightly more complicated when dealing with larger systems of equations, but when using only two equations it is as easy as the order condition. The rank condition is a necessary and sufficient condition, which means that if we can identify the equations using the rank condition we can be sure that the equation really is identified. The rank condition investigates whether two or more equations are linearly dependent on each other, which would be the case if the sum of two equations would equal a third equation in the model. If that is the case it is impossible to identify all structural parameters. The basic steps in this decision rule is best described by an example.

Example 12.3
Consider the following system of equations:

$$Y_1 = A_0 + A_1 Y_3 + A_2 X_1 + A_3 X_3 + U_1 \tag{12.20}$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + B_3 X_3 + U_2 \tag{12.21}$$

$$Y_3 = C_0 + C_1 Y_2 + A_2 X_1 + A_3 X_2 + U_3 \tag{12.22}$$

This system contains three endogenous variables (Y_1, Y_2, Y_3) and three exogenous variables (X_1, X_2, X_3), which means that we in total has six variables. The first step in checking the rank condition is to put up a matrix that for each equation mark which of the six variables that are included (marked with 1) and which that are excluded (marked with 0) from the equation. For our system we receive the following matrix:

Matrix for the rank condition

	Y_1	Y_2	Y_3	X_1	X_2	X_3
Equation 1	1	0	1	1	0	1
Equation 2	1	1	0	0	1	1
Equation 3	0	1	1	1	1	0

In order to check the rank condition for the first equation we have to proceed as follows: Delete the first row and collect the columns for those variables of the first equation that were marked with zero. For equation 1, Y_2 and X_2 was marked with zero, and if we collect those two columns we receive:

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

If this matrix contains less than M-1 rows or columns where all elements are zero, equation 1 will not be identified. M refers to the number of equation just as in the order condition, which means that M-1=2. Since we have two rows and two columns and none of them contains only zeros we conclude that equation 1 is identified.

For equation 2 we proceed in the same way. We delete the second row and collect those columns where the elements of the second row were marked with a zero. For equation 2 that was the case for Y_3 and X_1, which is to say that these two variables was not included in equation 2. The resulting matrix for this case then becomes:

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

It looks in the same way as for equation 1, which means that we have two rows and two columns that is not only zeros. The same procedure should be done for the third equation and if you do that you will see that it is identified as well.

When using larger systems it is quite possible that the order condition says that a particular equation is identified even though the rank condition says it is not. When that happens it might still be possible to generate estimates, but those estimates will not have any economic meaning since they will represent averages of those equations that are linear combinations of each other. Hence, you should not be content that you have received identified results just because the order condition says so and the econometric software generates results for you. When using systems of more than two equations you should also confirm the identification using the rank condition.

12.4 Estimation methods

Once we have confirmed that our model is identified we can proceed with the estimation of the parameters of the structural coefficients. In this section we will present two methods of estimation that can be used to estimate coefficients of a simultaneous equation system.

12.4.1 Indirect Least Squares (ILS)

When all the equations are exactly identified one can use the method of Indirect Least Square to estimate the coefficients of the structural equations. It is done by the following three steps:

1) Form the reduced form equations
2) Estimate the coefficients of the reduced form using OLS
3) Use the estimated coefficients of the reduced form to derive the structural coefficients.

Example 12.4 ***(ILS)***
Consider the following simple macro economic model:

$$Y_t = C_t + I_t \tag{12.23}$$

$$C_t = B_0 + B_1 Y_t + U_t \tag{12.24}$$

This model has two endogenous variables (Y_t and C_t) and one exogenous variable (I_t), and we would like to estimate the coefficients of the behavioral equation. Since one of the variables of the model is excluded from the consumption function it is identified according to the order condition. The two structural equations could be used to form the reduced form equations for consumption. If we do that we receive:

$$C_t = \pi_0 + \pi_1 I_t + V_t \tag{12.25}$$

(13.3) and (13.4) show how the reduced form coefficients are related to the structural coefficients. By using the estimated values of the reduced form coefficients we can solve for the structural coefficients. We have:

$$\pi_0 = \frac{B_0}{1 - B_1} \tag{12.26}$$

$$\pi_1 = \frac{B_1}{1 - B_1} \tag{12.27}$$

(12.26) and (12.27) can now be used to solve for B_0 and B_1. Since (12.27) is an equation with only one unknown we solve for B_1 first (remember that π_1 is an estimate and therefore a number in this expression). Once we receive the value of B_1 we can use it in (12.26) to solve for B_0. Hence we receive:

$$B_0 = \frac{\pi_0}{1 - \pi_1} \text{ and } B_1 = \frac{\pi_1}{1 - \pi_1}$$

In order to determine the standard errors for B_0 and B_1 we can use linear approximations to their expression based on the standard errors and covariance of the reduced form estimated coefficients. It can be shown that the corresponding variance for B_0 and B_1 is:

$$V(B_0) \approx a^2\sigma_0^2 + b^2\sigma_1^2 + 2ab\sigma_{01}$$

$$V(B_1) \approx b^2\sigma_1^2$$

with $a = \frac{1}{1-\pi_0}$ and $b = \frac{1}{(1-\pi_1)^2}$ and where σ_0^2 is the variance of π_0, σ_1^2 the variance for π_1 and σ_{12} the covariance between π_0 and π_1.

ILS will result in consistent estimates but will still be biased in small samples. When using larger systems with more variables and equations it is often burdensome to find the estimates, and in those cases the equations are often over identified, which means that ILS cannot be used. For that reason ILS is not used very often in practice. Instead a much more popular method called 2SLS is used.

12.4.2 Two Stage Least Squares (2SLS)

The procedure of 2SLS is a method that allows you to receive consistent estimates of the structural coefficient when the equations are exactly identified as well as over identified. However, the estimates will still be biased in small samples.

Consider the following model

$$Y_1 = A_0 + A_1Y_2 + A_3X_1 + A_4X_2 + U_1 \tag{12.28}$$

$$Y_2 = B_0 + B_1Y_1 + B_3X_3 + B_4X_4 + U_2 \tag{12.29}$$

This model has two endogenous variables and four exogenous variables. The first equation (12.28) contains four variables which means that from a total of six variables, two has been omitted. That means that it is over identified. The same can be said about the second equation (12.29), which means that the model is identified. Since both equations are over identified we cannot estimate the structural parameters using ILS, but instead we are forced to use 2SLS. We will now focus the discussion on the estimation of the first equation.

The basic steps of 2SLS applied for equation (12.28):

Step 1 Derive the reduced form equation for Y_2 and estimate the predicted value of Y_2 $(\hat{Y}_2)$ on the reduced form using OLS.

$$\hat{Y}_2 = \hat{\pi}_0 + \hat{\pi}_1 X_1 + \hat{\pi}_2 X_2 + \hat{\pi}_3 X_3 + \hat{\pi}_4 X_4$$

Step 2 Replace Y_2 in equation (12.28) with its predicted value from the reduced form and estimate the coefficient of the model using OLS.

$$Y_1 = A_0 + A_1 \hat{Y}_2 + A_3 X_1 + A_4 X_2 + U_1$$

If these two steps are applied we will receive consistent estimates of the parameters in (12.28). That is, since we replace the endogenous variable with its predicted value, it is no longer correlated with the residual term. Hence, the problem is solved. Remember that $Y_2 = \hat{Y}_2 + V_2$ which implies that the stochastic variable Y_2, consist of two parts, one that is a linear combination of the exogenous (predetermined) variables and one random part. A group of exogenous variables are by necessity uncorrelated with the random term.

Observe that X_1 and X_2 both appear in the specification of Y_1 and $\hat{Y}_2$, which means that there will be a correlation between the explanatory variables X_1 and X_2 and $\hat{Y}_2$. This correlation will not be perfect unless X_3 and X_4 also is included in the structural model of the first equation and is therefore nothing to worry about. But if that happens, the equation would not pass the order condition for identification.

There is one additional complication to be aware of when working with 2SLS. When the predicted value is included in the specification, the variance of the error term will not be correct. To see this we will consider a simplified version of a model to make it clear where the problem appear. Consider the following equation:

$$Y_{1i} = B_1 Y_{2i} + U_{1i} \tag{12.30}$$

In order to receive consistent estimates of B_1 we replace Y_2 with its predicted value and estimate the parameters of following regression model using OLS:

$$Y_{1i} = B_1 \hat{Y}_{2i} + (U_{1i} + B_1 U_{2i}) = B_1 \hat{Y}_{2i} + V_1 \tag{12.31}$$

The estimator of the slope coefficient is therefore given by the following expression:

$$b_1 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)Y_{1i}}{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)^2} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)(B_1Y_{2i} + U_{1i})}{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)^2} = B_1 + \frac{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)U_{1i}}{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)^2} \tag{12.32}$$

We have concluded that this form of the estimator is consistent but not unbiased. Since it is consistent we need to compare it with its asymptotic variance, that is, the formula of the variance when the number of observation is very large (has gone to infinity). It can be shown that the asymptotic variance of this sample estimator is given by the following expression:

$$V(b_1) = \frac{\sigma_U^2}{\sum_{i=1}^{n}\left(\hat{Y}_{2i} - \bar{\hat{Y}}_2\right)^2} \tag{12.33}$$

This is good, because it is very similar to the variance given by the standard OLS. So what is the problem? The problem is related to the estimated variance of the error term. When running the regression using (12.31) our estimated residual would be given by:

$$\hat{\sigma}_V^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{1i} - b_1\hat{Y}_{2i}\right)^2 \tag{12.34}$$

Whereas the estimated residual should be given by the following expression

$$\hat{\sigma}_U^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{1i} - b_1Y_{2i}\right)^2 \tag{12.34}$$

(12.34) is based on the observed variable Y_2 multiplied with the sample estimator b_1 given by (12.32), rather than the predicted version of the variable. Hence in order to receive consistent estimates of the standard errors, one has to use (12.34). When using commercial software with routines for 2SLS they automatically make the correction. But if you run 2SLS in two steps, as described above, you need to correct the standard errors, before you can perform and hypothesis testing.

In sum, the 2SLS has the following properties:

- It generates biased but consistent estimates
- The distribution of the estimators are normally distributed only in large samples
- The variance is biased but consistent when using (12.34)

A. Statistical tables

Table A1

Area below the standard normal distribution: $P(Z \le z)$

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999
3.1	0.99903	0.99906	0.9991	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.9994	0.99942	0.99944	0.99946	0.99948	0.9995
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.9996	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.9997	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

Example: Find the probability that Z is less then 1.33: $P(Z \le 1.33) = 0.90824$

Table A2

Right tail critical values for the t-distribution

	Probability						
Degrees of	0.25	0.10	0.05	0.025	0.01	0.005	0.001
freedom	0.50	0.20	0.10	0.050	0.02	0.010	0.002
1	1.000001	3.077685	6.313749	12.70615	31.82096	63.6559	318.2888
2	0.816497	1.885619	2.919987	4.302656	6.964547	9.924988	22.32846
3	0.764892	1.637745	2.353363	3.182449	4.540707	5.840848	10.21428
4	0.740697	1.533206	2.131846	2.776451	3.746936	4.60408	7.17293
5	0.726687	1.475885	2.015049	2.570578	3.36493	4.032117	5.893526
6	0.717558	1.439755	1.943181	2.446914	3.142668	3.707428	5.207548
7	0.711142	1.414924	1.894578	2.364623	2.997949	3.499481	4.785252
8	0.706386	1.396816	1.859548	2.306006	2.896468	3.355381	4.500762
9	0.702722	1.383029	1.833114	2.262159	2.821434	3.249843	4.29689
10	0.699812	1.372184	1.812462	2.228139	2.763772	3.169262	4.143658
11	0.697445	1.36343	1.795884	2.200986	2.718079	3.105815	4.024769
12	0.695483	1.356218	1.782287	2.178813	2.68099	3.054538	3.929599
13	0.69383	1.350172	1.770932	2.160368	2.650304	3.012283	3.852037
14	0.692417	1.345031	1.761309	2.144789	2.624492	2.976849	3.787427
15	0.691197	1.340605	1.753051	2.131451	2.602483	2.946726	3.732857
16	0.690133	1.336757	1.745884	2.119905	2.583492	2.920788	3.686146
17	0.689195	1.333379	1.739606	2.109819	2.56694	2.898232	3.645764
18	0.688364	1.330391	1.734063	2.100924	2.552379	2.878442	3.610476
19	0.687621	1.327728	1.729131	2.093025	2.539482	2.860943	3.579335
20	0.686954	1.325341	1.724718	2.085962	2.527977	2.845336	3.551831
21	0.686352	1.323187	1.720744	2.079614	2.517645	2.831366	3.527093
22	0.685805	1.321237	1.717144	2.073875	2.508323	2.818761	3.504974
23	0.685307	1.319461	1.71387	2.068655	2.499874	2.807337	3.484965
24	0.68485	1.317835	1.710882	2.063898	2.492161	2.796951	3.466776
25	0.68443	1.316346	1.70814	2.059537	2.485103	2.787438	3.450186
26	0.684043	1.314972	1.705616	2.055531	2.478628	2.778725	3.43498
27	0.683685	1.313704	1.703288	2.051829	2.472661	2.770685	3.42101
28	0.683353	1.312526	1.70113	2.048409	2.467141	2.763263	3.408204
29	0.683044	1.311435	1.699127	2.045231	2.46202	2.756387	3.396271
30	0.682755	1.310416	1.69726	2.04227	2.457264	2.749985	3.385212
35	0.681564	1.306212	1.689573	2.03011	2.437719	2.723809	3.340028
40	0.680673	1.303076	1.683852	2.021075	2.423258	2.704455	3.306923
45	0.679981	1.30065	1.679427	2.014103	2.412116	2.689594	3.281457
50	0.679428	1.298713	1.675905	2.00856	2.403267	2.677789	3.261375
60	0.678601	1.295821	1.670649	2.000297	2.390116	2.660272	3.231689
80	0.677569	1.292224	1.664125	1.990065	2.373872	2.638699	3.195237
100	0.676951	1.290075	1.660235	1.983972	2.364213	2.625893	3.173773
∞	0.67449	1.281551	1.644853	1.959966	2.326351	2.575835	3.090245

Note: The smaller value at the head of each column is the area in one tail, the larger value is the area in both tails.

Table A3

Right tail critical value of the Chi-Square distribution

	Probability				
Degrees of freedom	0.100	0.050	0.025	0.010	0.005
1	2.705541	3.841455	5.023903	6.634891	7.8794
2	4.605176	5.991476	7.377779	9.210351	10.59653
3	6.251394	7.814725	9.348404	11.34488	12.83807
4	7.779434	9.487728	11.14326	13.2767	14.86017
5	9.236349	11.07048	12.83249	15.08632	16.74965
6	10.64464	12.59158	14.44935	16.81187	18.54751
7	12.01703	14.06713	16.01277	18.47532	20.27774
8	13.36156	15.50731	17.53454	20.09016	21.95486
9	14.68366	16.91896	19.02278	21.66605	23.58927
10	15.98717	18.30703	20.4832	23.20929	25.18805
11	17.27501	19.67515	21.92002	24.72502	26.75686
12	18.54934	21.02606	23.33666	26.21696	28.29966
13	19.81193	22.36203	24.73558	27.68818	29.81932
14	21.06414	23.68478	26.11893	29.14116	31.31943
15	22.30712	24.9958	27.48836	30.57795	32.80149
16	23.54182	26.29622	28.84532	31.99986	34.26705
17	24.76903	27.5871	30.19098	33.40872	35.71838
18	25.98942	28.86932	31.52641	34.80524	37.15639
19	27.20356	30.14351	32.85234	36.19077	38.58212
20	28.41197	31.41042	34.16958	37.56627	39.99686
21	29.61509	32.67056	35.47886	38.93223	41.40094
22	30.81329	33.92446	36.78068	40.28945	42.79566
23	32.00689	35.17246	38.07561	41.63833	44.18139
24	33.19624	36.41503	39.36406	42.97978	45.55836
25	34.38158	37.65249	40.6465	44.31401	46.92797
26	35.56316	38.88513	41.92314	45.64164	48.28978
27	36.74123	40.11327	43.19452	46.96284	49.64504
28	37.91591	41.33715	44.46079	48.27817	50.99356
29	39.08748	42.55695	45.72228	49.58783	52.3355
30	40.25602	43.77295	46.97922	50.89218	53.67187
35	46.05877	49.80183	53.20331	57.34199	60.27459
40	51.80504	55.75849	59.34168	63.69077	66.76605
45	57.50529	61.65622	65.41013	69.9569	73.16604
50	63.16711	67.50481	71.42019	76.1538	79.48984
60	74.397	79.08195	83.29771	88.37943	91.95181
80	96.5782	101.8795	106.6285	112.3288	116.3209
100	118.498	124.3421	129.5613	135.8069	140.1697

Table A4

Right tail critical for the F-distribution: 5 percent level

*n**m*	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95	248.02	249.26	250.10	251.14	252.20	253.25	254.32
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32	1.22	1.00

Note: *m* = degrees of freedom for the numerator
n = degrees of freedom for the denominator