PUBLISHED BY
bookboon

Thomas Andren

Econometrics – Part II

Thomas Andren

Econometrics

Part II

Econometrics – Part II
1st edition

ISBN 978-87-7681-316-1

Contents

1 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 chapter in book 1 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:

$$b_0 \sim N\left(B_0, \frac{\sigma^2 \sum X_{1i}^2}{\sum \left(X_{1i} - \overline{X}_1\right)^2}\right) \tag{1.1}$$

$$b_1 \sim N\left(B_1, \frac{\sigma^2}{\sum \left(X_{1i} - \overline{X}_1\right)^2}\right) \tag{1.2}$$

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.

1.1 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 1.1

Assume the following population regression equation:

$$Y = B_0 + B_1 X + U \tag{1.3}$$

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

$$\begin{aligned} E[Y \mid X] &= 4.233 + 0.469X \\ &\quad (3.26)\ \ (0.213) \end{aligned} \tag{1.4}$$

The regression results in (1.4) 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:

$$\begin{aligned} &H_0 : B_1 = 0 \\ &H_1 : B_1 \neq 0 \end{aligned} \tag{1.5}$$

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:

$$\text{Test function:} \quad t = \frac{b_1 - B_1}{se(b_1)} \sim t_{n-k} \tag{1.6}$$

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:

$$\text{Test value:} \qquad t = \frac{0.469 - 0}{0.213} = 2.2 \tag{1.7}$$

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:

$$\text{Critical value: } t_c = 1.96 \tag{1.8}$$

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.

1.2 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 (1.5) 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:

$$P(-1.96 \leq t \leq 1.96) = 95\% \tag{1.9}$$

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 (1.6) into (1.9). 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:

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

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.

1.2.1 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.

	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

Table 1.1 Regression output from Excel

In Table 1.1 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.

1.3 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.

$$P(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

The type II error

The probability to accept a false null hypothesis

$$P(\text{Accept } H_0 \mid H_0 \text{ is false}) = \beta$$

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

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

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

Table 1.2 Errors in hypothesis testing

Example 1.2

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 1.2. Let us use the regression results from Example 1.1 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 1.1 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:

$$\left[-1.96 \times se(b_1), 1.96 \times se(b_1)\right] = \left[-0.41748, 0.41748\right] \tag{1.10}$$

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:

$$b_1 \sim N\left(0.25, V(b_1)\right) \tag{1.11}$$

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 (1.10). That is, we have to calculate the following probability using the distribution given by (1.11):

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

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:

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

$$0.783 - 0.0009 = 0.7821$$

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 (1.10) 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.

1.4 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:

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

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:

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

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:

$$\sigma_f^2 = E\left[Y_{T+1} - \hat{Y}_{T+1}\right]^2 = \sigma^2 \left[1 + \frac{1}{T} + \frac{\left(X_{T+1} - \overline{X}\right)}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}\right] \tag{1.12}$$

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:

$$\sigma_f^2 = \sigma^2\left[1+\frac{1}{T}+\frac{\left(X_{T+1}-\overline{X}\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}+\frac{\overline{X}_{T+1}\sigma_\varepsilon^2}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}\right]+B_1^2\sigma_\varepsilon^2 \tag{1.13}$$

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 (1.12) 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 (1.12) or (1.13) 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

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

2 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.

2.1 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:

$$Y_i - \overline{Y} = Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i = \underbrace{\left(\hat{Y}_i - \overline{Y}\right)}_{\text{Explained}} + \underbrace{\left(Y_i - \hat{Y}_i\right)}_{\text{Unexplained}} \tag{2.1}$$

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 (2.1) 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:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}_{TSS}=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}_{ESS}+\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{RSS} \tag{2.2}$$

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:

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

Example 2.1

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:

$$R^2 = \frac{ESS}{TSS} = \frac{b_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \left(b_1 \frac{S_X}{S_Y}\right)^2 \tag{2.3}$$

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.

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X^2} \times \frac{S_Y}{S_Y} = \frac{S_{XY}}{S_X S_Y} \times \frac{S_Y}{S_X} = r \frac{S_Y}{S_X} \tag{2.4}$$

where S_{XY} represents the sample covariance between X and Y, and r the sample correlation coefficient for X and Y. Hence, substituting (2.4) into (2.3) shows the relation between the sample correlation coefficient and the coefficient of determination.

$$R^2 = \left(b_1 \frac{S_X}{S_Y}\right)^2 = \left(r \times \frac{S_X}{S_Y} \times \frac{S_Y}{S_X}\right)^2 = (r)^2 \tag{2.5}$$

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

$$r = \sqrt{R^2} \tag{2.6}$$

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 (2.3)–(2.6) 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.

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

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.

2.2 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:

$$\overline{R}^2 = 1 - \frac{RSS/(n-2)}{TSS/(n-1)} = 1 - \frac{Var(e)}{Var(Y)} \tag{2.7}$$

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 (2.7) 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 (2.7) 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.

2.3 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:

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 2, book 2and 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:

$$F = \frac{\text{Additional variance explained by } X}{\text{Unexplained variance}} \tag{2.8}$$

In the numerator of equation (2.8) 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:

$$F = \frac{ESS/1}{RSS/(n-2)} \sim F_{(1,n-2)} \tag{2.9}$$

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

Example 2.2

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 (2.9) 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:

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

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.

3 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 next chapter. This chapter will focus on the differences between the simple and the multiple-regression model and extend the concepts from the previous chapters.

3.1 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:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{3.1}$$

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 3.1

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:

$$P = B_0 + B_1 A + B_2 K + U \tag{3.2}$$

The model offers the following two marginal effects:

$$\frac{\partial P}{\partial A} = B_1 \tag{3.3}$$

$$\frac{\partial P}{\partial K} = B_2 \tag{3.4}$$

The first marginal effect (3.3) 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 (3.4) 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:

$$P = B_0 + B_1 A + B_2 A^2 + B_3 K + B_4 K^2 + B_5 A \times K + U \tag{3.5}$$

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

$$\frac{\partial P}{\partial A} = B_1 + 2B_2 A + B_5 K \tag{3.6}$$

$$\frac{\partial P}{\partial K} = B_3 + 2B_4 K + B_5 A \tag{3.7}$$

Equation (3.6) 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 (3.6) and (3.7) consist of three parameter estimates, which individually can be interpreted.

Focusing on (3.6) 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.

3.2 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 (3.2) are given by the following expressions:

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 \tag{3.8}$$

$$b_1 = \frac{S_{Y1} - r_{12}r_{Y2}S_Y S_1}{S_1^2(1-r_{12}^2)} \tag{3.9}$$

$$b_2 = \frac{S_{Y2} - r_{12}r_{Y1}S_Y S_2}{S_2^2(1-r_{12}^2)} \tag{3.10}$$

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 (3.9) we have that

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

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 3. Equation (3.8)–(3.10) 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.

3.3 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:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + U \tag{3.11}$$

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*.

3.3.1 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: $$Y = B_0 + B_3 X_3 + U \tag{3.12}$$

Model according to the alternative hypothesis: $$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + U \tag{3.13}$$

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 (3.12) you should think of it as a restricted version of the full model given by (3.13) since two of the parameters are forced to zero. In (3.13) 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* $\left(RSS_R\right)$ received from (3.12) and an Unrestricted *RSS* $\left(RSS_U\right)$ received from (3.13). 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:

$$F = \frac{\left(RSS_R - RSS_U\right)/df_1}{RSS_U/df_2} \sim F_{(\alpha,df_1,df_2)} \tag{3.14}$$

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 $k_2\text{-}k_1=2$.

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 (3.14) 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 3.2

Consider the two specifications given by (3.12) and (3.13), 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 3.1.

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

Table 3.1 Summary results from the two regressions

Using the information in Table 3.1 we may calculate the test value for our test.

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

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.

3.3.2 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: $Y = B_0 + U$ (3.15)

Model according to the alternative hypothesis: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + U$ (3.16)

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 3.3

Assume that we have access to a sample of 1000 observations and that we would like to estimate the parameters in (3.16), and test the over all significance of the model. Running the regression using our sample we received the following ANOVA table:

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

Table 3.2 ANOVA table

Using the information from Table 3.2 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.

4 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.

4.1 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.

4.1.1 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:

$$Y = B_0 + B_1 X + U \tag{4.1}$$

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

$$\frac{dY}{dX} = B_1 \tag{4.2}$$

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:

$$e = \frac{dY}{dX}\frac{X}{Y} = B_1 \frac{X}{Y} \tag{4.3}$$

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 4.1

Calculate the marginal effect and the elasticity using the regression results in Table 4.1 received from a sample of data using model (4.1).

	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

Table 4.1 Regression results from a model with a linear specification

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.

4.1.2 The log-linear specification

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

$$\ln Y = B_0 + B_1 X + U \tag{4.4}$$

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

$$\ln(w_s) = s\ln(1+r) + \ln(w_0) = B_0 + B_1 s \tag{4.5}$$

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 (4.4). 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:

$$\frac{d\ln(w_s)}{ds} = \frac{1}{w_s}\frac{dw_s}{ds} = \frac{dw_s / w_s}{ds} = B_1 \tag{4.6}$$

Expression (4.6) 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 (4.6) we see that the marginal effect is given by:

$$\frac{dw_s}{ds} = B_1 \times w_s \tag{4.7}$$

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 (4.7) 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.

4.1.3 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

$$Y = B_0 + B_1 \ln X + U \tag{4.8}$$

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

$$\frac{dY}{dX} = \frac{1}{X}B_1 \quad <=> \quad \frac{dY}{dX / X} = B_1 \tag{4.9}$$

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.

4.1.4 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:

$$Q(L,K) = AL^{B_1}K^{B_2} \tag{4.10}$$

(4.10) 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:

$$\ln Q = \ln A + B_1 \ln L + B_2 \ln K + U = B_0 + B_1 \ln L + B_2 \ln K + U \tag{4.11}$$

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 (4.11). 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:

$$\frac{\partial \ln Q}{\partial \ln L} = \frac{\partial Q / Q}{\partial L / L} = \frac{\partial Q}{\partial L}\frac{L}{Q} = B_1 \tag{4.12}$$

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 (4.12) and equals:

$$\frac{\partial Q}{\partial L} = B_1 \frac{Q}{L} \tag{4.13}$$

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

4.2 Omission of a relevant variable

In chapter 3, book 1 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: $$Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{4.14}$$

The estimated model: $$Y = b_0 + b_1 X_1 + e \tag{4.15}$$

From chapter 3, book 1 we know that the sample estimator for the slope coefficient in the simple regression model is given by:

$$b_1 = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)Y_i}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} = \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)\left(B_0 + B_1 X_{1i} + B_2 X_{2i} + U_i\right)}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} \tag{4.16}$$

which may be rewritten as

$$b_1 = B_0 \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} + B_1 \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)X_{1i}}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} + B_2 \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)X_{2i}}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} + \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)U_i}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} \tag{4.17}$$

Simplify and take the expectation of the estimator:

$$E[b_1] = B_1 + B_2 \frac{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)X_{2i}}{\sum_{i=1}^{n}\left(X_{1i} - \overline{X}_1\right)^2} = B_1 + B_2 \frac{Cov(X_1, X_2)}{Var(X_1)} \tag{4.18}$$

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 (4.18) 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 (4.18).

4.3 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: $$Y = B_0 + B_1 X_1 + U \tag{4.19}$$

The estimated model: $$Y = b_0 + b_1 X_1 + b_2 X_2 + e \tag{4.20}$$

The estimated model (4.20) 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:

$$b_1 = \frac{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)Y_i - \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)Y_i}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2 - \left(\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)\right)^2} \tag{4.21}$$

Substitute the (4.19) 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**.

4.4 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:

$$Y^* = Y + \varepsilon \tag{4.22}$$

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 (4.22) with Y:

$$\begin{aligned} Y &= B_0 + B_1 X + U \\ Y^* - \varepsilon &= B_0 + B_1 X + U \\ Y^* &= B_0 + B_1 X + (U + \varepsilon) = B_0 + B_1 X + U^* \end{aligned} \tag{4.23}$$

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:

$$X^* = X + \varepsilon \tag{4.24}$$

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

$$\begin{aligned} Y &= B_0 + B_1(X^* - \varepsilon) + U \\ Y &= B_0 + B_1 X^* + (U - B_1\varepsilon) = B_0 + B_1 X^* + U^* \end{aligned} \tag{4.25}$$

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.

5 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:

$$D = \begin{cases} 1 & \text{if a man} \\ 0 & \text{otherwise (if a woman)} \end{cases} \tag{5.1}$$

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 (5.1) which is standard.

5.1 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 (5.1) we can form the following model with two explanatory variables.

$$Y = B_0 + B_1 D + B_2 X + U \tag{5.2}$$

As can be seen from (5.1) D takes only two values. If we form the conditional expectation with respect to the two categories of D we receive:

$$E[Y \mid D = 1, X] = B_0 + B_1 + B_2 X \tag{5.3}$$
$$E[Y \mid D = 0, X] = B_0 + B_2 X \tag{5.4}$$

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 (5.3) 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 (5.4) 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:

$$E[Y \mid D=1, X] - E[Y \mid D=0, X] = B_1 \tag{5.5}$$

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 (5.5) 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 5.1

Assume the following regression result from a model given by (5.2) 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:

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

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:

$$\ln Y = B_0 + B_1 D + B_2 X + U \tag{5.6}$$

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:

$$\hat{Y} = e^{(B_0 + B_1 D + B_2 X + \sigma_U^2 / 2)} \tag{5.7}$$

where σ_U^2 represents the population variance of the error term. Hence, in order to receive the conditional expectation given by (5.7) 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 (5.7) instead. Doing that we receive:

Marginal effect:

$$\frac{E[Y \mid D = 1, X] - E[Y \mid D = 0, X]}{E[Y \mid D = 0, X]} = \frac{e^{\left(B_0 + B_1 + B_2 X + \sigma_U^2 / 2\right)} - e^{\left(B_0 + B_2 X + \sigma_U^2 / 2\right)}}{e^{\left(B_0 + B_2 X + \sigma_U^2 / 2\right)}} = e^{B_1} - 1 \tag{5.8}$$

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: $$V\left(e^{b_1}-1\right)=V\left(e^{b_1}\right)=\sigma_{b_1}^2\times\left(e^{b_1}\right)^2 \tag{5.9}$$

Example 5.2

Assume the following regression results, from a model given by (5.6), 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:

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

Since we are interested in the marginal effect of D on Y, we have to calculate it using the regression results. By (5.8) and (5.9) 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 (5.8). 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 (5.8), and is therefore often used directly as such.

Observe that

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

therefore researcher often use b_1 directly instead of the calculated value given by (5.8).

5.2 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:

$$\begin{aligned} \ln Y &= B_0 + (B_1 + B_2 D)X + U \\ &= B_0 + B_1 X + B_2 (DX) + U \end{aligned} \tag{5.10}$$

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 5.3

Use the same data set as in Example 5.2 and estimate the coefficients in (5.10). The results are presented below with standard errors within parenthesis:

$$\begin{matrix}\ln \hat{Y} = 4.11 + 0.024X + 0.014DX \\ \quad (0.031)(0.003) \quad (0.001)\end{matrix} \tag{5.11}$$

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.

5.2.1 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:

$$\ln Y = B_0 + B_1 X + B_2 D + B_3 (DX) + U \tag{5.12}$$

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 (5.10), but not otherwise.

Example 5.4

Extend the specification of (5.10) by including D separately. That is, estimate the parameters of the model given by (5.12) and interpret the results. Doing that, we received the following results, with standard errors within parenthesis.

$$\begin{matrix}\ln \hat{Y} = 4.006 + 0.033X + 0.210D - 0.002DX \\ \quad (0.045) \quad (0.004) \quad (0.062) \quad (0.005)\end{matrix} \tag{5.12}$$

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 5.3 and 5.4 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 5.3 we did not include D even though it was relevant. In last chapter 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.

5.3 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:

$$D = \begin{cases} 0 & \text{Primary schooling} \\ 1 & \text{Secondary schooling} \\ 2 & \text{Post secondary schooling} \end{cases} \tag{5.13}$$

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:

$$\ln Y = B_0 + B_1 D_1 + B_2 D_2 + B_3 D_3 + B_4 X + U \tag{5.14}$$

It is a mathematical impossibility to estimate the parameters in (5.14) 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:

$$\ln Y = B_0 + B_2 D_2 + B_3 D_3 + B_4 X + U \tag{5.15}$$

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 (5.8).

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

$$\ln Y = C_1 D_1 + C_2 D_2 + C_3 D_3 + B_4 X + U \tag{5.16}$$

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 5.5

Estimate the parameters of (5.15) and (5.16) and compare and interpret the results.

Specification I: (5.15)
$$\underset{\quad}{\ln \hat{Y} = 3.929 + 0.154 D_2 + 0.295 D_3 + 0.009 X} \tag{5.17}$$
$$(0.043)\quad (0.024)\quad (0.022)\quad (0.001)$$

Specification II: (5.16)
$$\underset{\quad}{\ln \hat{Y} = 3.929 D_1 + 4.083 D_2 + 4.224 D_3 + 0.009 X} \tag{5.18}$$
$$(0.043)\quad (0.034)\quad (0.036)\quad (0.001)$$

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.

5.4 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 5.1 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:

$$Y = A + BX + U \tag{5.19}$$

In order to transform (5.19) 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:

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

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

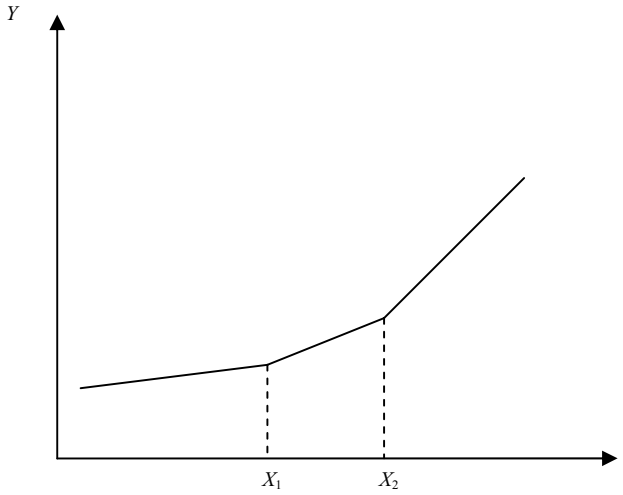

Figure 5.1 Piecewise linear regression

Next re-specify the intercept and the slope coefficient in (5.19) in the following way:

$$A = A_0 + A_1 D_1 + A_2 D_2 \tag{5.20}$$

$$B = B_0 + B_1 D_1 + B_2 D_2 \tag{5.21}$$

Substitute (5.20) and (5.21) into (5.19) 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:

$$Y = A_0 + A_1 D_1 + A_2 D_2 + B_0 X + B_1 (D_1 X) + B_2 (D_2 X) + U \tag{5.22}$$

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

When $X < X_1$ $\hat{Y} = a_0 + b_0 X$

When $X_1 \le X \le X_2$ $\hat{Y} = (a_0 + a_1) + (b_0 + b_1) X$

When $X > X_2$ $\hat{Y} = (a_0 + a_2) + (b_0 + b_2) X$

5.5 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:

$$\ln Y = B_0 + B_1 X_1 + B_2 X_2 + U \tag{5.23}$$

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:

$$H_0 : B_{1m} = B_{1w}, B_{2m} = B_{2w}$$

$$H_1 : B_{1m} \neq B_{1w} \quad and/or \quad B_{2m} \neq B_{2w}$$

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:

$$\ln Y = B_0 + B_1 D + B_2 X_1 + B_3 (X_1 D) + B_4 X_2 + B_5 (X_2 D) + U \tag{5.24}$$

Equation (5.24) will be representing the unrestricted model, where men and women are allowed to have different coefficients, and equation (5.23) 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:

$$F = \frac{(RSS_R - RSS_U)/df_1}{RSS_U/df_2} \sim F_{(J,n-k)} \tag{5.25}$$

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 5.6

Assume that would like to know if the coefficients of equation (5.23) 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 (5.23) and one unrestricted model given by (5.24). Using a sample of 1483 randomly selected individuals we received the following 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

Table 5.1 Regression results

Using the results given in Table 5.1 we may calculate the test value using (5.25). 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.