4 Hypothesis testing in the multiple regression model

Ezequiel Uriel
Universidad de Valencia
Version: 09-2013

4.1 Hypothesis testing: an overview

Before testing hypotheses in the multiple regression model, we are going to offer a general overview on hypothesis testing.

Hypothesis testing allows us to carry out inferences about population parameters using data from a sample. In order to test a hypothesis in statistics, we must perform the following steps:

1) Formulate a null hypothesis and an alternative hypothesis on population parameters.

2) Build a statistic to test the hypothesis made.

3) Define a decision rule to reject or not to reject the null hypothesis.

Next, we will examine each one of these steps.

4.1.1 Formulation of the null hypothesis and the alternative hypothesis

Before establishing how to formulate the null and alternative hypothesis, let us make the distinction between *simple* hypotheses and *composite* hypotheses. The hypotheses that are made through one or more equalities are called simple hypotheses. The hypotheses are called composite when they are formulated using the operators "inequality", "greater than" and "smaller than".

It is very important to remark that hypothesis testing is always about *population* parameters. Hypothesis testing implies making a decision, on the basis of sample data, on whether to reject that certain restrictions are satisfied by the basic assumed model. The restrictions we are going to test are known as the *null hypothesis*, denoted by H_0. Thus, null hypothesis is a statement on population parameters.

Although it is possible to make composite null hypotheses, in the context of the regression model the null hypothesis is always a simple hypothesis. That is to say, in order to formulate a null hypothesis, which shall be called H_0, we will always use the operator "equality". Each equality implies a restriction on the parameters of the model. Let us look at a few examples of null hypotheses concerning the regression model:

a) $H_0 : \beta_1=0$

b) $H_0 : \beta_1+ \beta_2=0$

c) $H_0 : \beta_1=\beta_2=0$

d) $H_0 : \beta_2+\beta_3=1$

We will also define an *alternative* hypothesis, denoted by H_1, which will be our conclusion if the experimental test indicates that H_0 is false.

Although the alternative hypotheses can be simple or composite, in the regression model we will always take a composite hypothesis as an alternative hypothesis. This hypothesis, which shall be called H_1, is formulated using the operator "inequality" in most cases. Thus, for example, given the H_0:

$$H_0 : \beta_j = 1 \tag{4-1}$$

we can formulate the following H_1:

$$H_1 : \beta_j \neq 1 \tag{4-2}$$

which is a "two side alternative" hypothesis.

The following hypotheses are called "one side alternative" hypotheses

$$H_1 : \beta_j < 1 \tag{4-3}$$

$$H_1 : \beta_j > 1 \tag{4-4}$$

4.1.2 Test statistic

A *test statistic* is a function of a random sample, and is therefore a random variable. When we compute the statistic for a given sample, we obtain an outcome of the test statistic. In order to perform a statistical test we should know the distribution of the test statistic under the null hypothesis. This distribution depends largely on the assumptions made in the model. If the specification of the model includes the assumption of normality, then the appropriate statistical distribution is the normal distribution or any of the distributions associated with it, such as the Chi-square, Student's *t*, or Snedecor's *F*.

Table 4.1 shows some distributions, which are appropriate in different situations, under the assumption of normality of the disturbances.

TABLE 4.1. Some distributions used in hypothesis testing.

	1 restriction	*1 or more restrictions*
Known σ^2	*N*	*Chi-square*
Unknown σ^2	Student's *t*	Snedecor's *F*

The statistic used for the test is built taking into account the H_0 and the sample data. In practice, as σ^2 is always unknown, we will use the distributions *t* and *F*.

4.1.3 Decision rule

We are going to look at two approaches for hypothesis testing: the classical approach and an alternative one based on *p*-values. But before seeing how to apply the decision rule, we shall examine the types of mistakes that can be made in testing hypothesis.

Types of errors in hypothesis testing

In hypothesis testing, we *can* make two kinds of errors: *Type I error* and *Type II error*.

Type I error

We can reject H_0 when it is in fact true. This is called *Type I error*. Generally, we define the *significance level* (α) of a test as the probability of making a *Type I error*. Symbolically,

$$\alpha = \Pr(\textit{Reject } H_0 \mid H_0) \qquad (4\text{-}5)$$

In other words, the significance level is the probability of rejecting H_0 given that H_0 is true. Hypothesis testing rules are constructed making the probability of a *Type I error* fairly small. Common values for α are 0.10, 0.05 and 0.01, although sometimes 0.001 is also used.

After we have made the decision of whether or not to reject H_0, we have either decided correctly or we have made an error. We shall never know with certainty whether an error was made. However, we can compute the *probability* of making either a *Type I error* or a *Type II error*.

Type II error

We can fail to reject H_0 when it is actually false. This is called *Type II error*.

$$\beta = \Pr(\textit{No reject } H_0 \mid H_1) \qquad (4\text{-}6)$$

In words, β is the probability of not rejecting H_0 given that H_1 is true.

It is not possible to minimize both types of error simultaneously. In practice, what we do is select a low significance level.

Classical approach: Implementation of the decision rule

The classical approach implies the following steps:

a) *Choosing* α. Classical hypothesis testing requires that we initially specify a *significance level* for the test. When we specify a value for α, we are essentially quantifying our tolerance for a *Type I error*. If α=0.05, then the researcher is willing to falsely reject H_0 5% of the time.

b) *Obtaining c*, the *critical value*, using statistical tables. The value *c* is determined by α.

The critical value (c) for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected.

c) *Comparing the outcome of the test* statistic, s, with c, H_0 is either rejected or not for a given α.

The rejection region (*RR*), delimited by the critical value(s), is a set of values of the test statistic for which the null hypothesis is rejected. (See figure 4.1). That is, the sample space for the test statistic is partitioned into two regions; one region (the rejection region) will lead us to reject the null hypothesis H_0, while the other will lead us not to reject the null hypothesis. Therefore, if the observed value of the test statistic S is in the critical region, we conclude by *rejecting* H_0; if it is not in the rejection region then we conclude by *not rejecting* H_0 or *failing to reject* H_0.

Symbolically,

$$\begin{array}{lll} \text{If} & s \geq c & \text{reject} \quad H_0 \\ \text{If} & s < c & \text{not reject} \ H_0 \end{array} \tag{4-7}$$

If the null hypothesis is rejected with the evidence of the sample, this is a *strong* conclusion. However, the acceptance of the null hypothesis is a *weak* conclusion because we do not know what the probability is of not rejecting the null hypothesis when it should be rejected. That is to say, we do not know the probability of making a type II error. Therefore, instead of using the expression of accepting the null hypothesis, it is more correct to say *fail to reject* the null hypothesis, or *not reject*, since what really happens is that we do not have enough empirical evidence to reject the null hypothesis.

In the process of hypothesis testing, the most subjective part is the *a priori* determination of the significance level. What criteria can be used to determine it? In general, this is an arbitrary decision, though, as we have said, the 1%, 5% and 10% levels for α are the most used in practice. Sometimes the testing is made conditional on several significance levels.

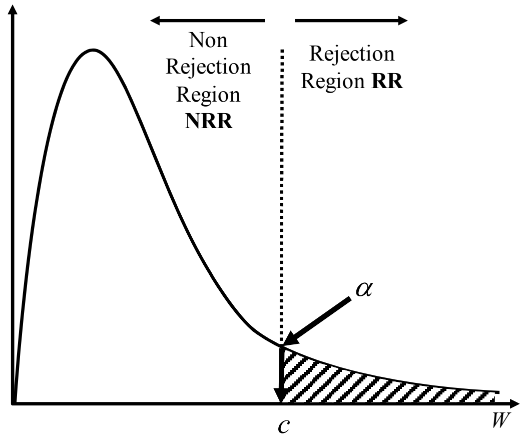


FIGURE 4.1. Hypothesis testing: classical approach.

An alternative approach: p-value

With the use of computers, hypothesis testing can be contemplated from a more rational perspective. Computer programs typically offer, together with the test statistic, a probability. This probability, which is called p-value (i.e., probability value), is also known as the critical or exact level of significance or the exact probability of making a

Type I error. More technically, the *p* value is defined as the lowest significance level at which a null hypothesis can be rejected.

Once the *p*-value has been determined, we know that the null hypothesis is rejected for any $\alpha \geq p$-value, while the null hypothesis is not rejected when $\alpha < p$-value. Therefore, the *p*-value is an indicator of the level of admissibility of the null hypothesis: the higher the *p*-value, the more confidence we can have in the null hypothesis. The use of the *p*-value turns hypothesis testing around. Thus, instead of fixing *a priori* the significance level, the *p*-value is calculated to allow us to determine the significance levels of those in which the null hypothesis is rejected.

In the following sections, we will see the use of *p* value in hypothesis testing put into practice.

4.2 Testing hypotheses using the *t* test

4.2.1 Test of a single parameter

The t test

Under the *CLM* assumptions 1 through 9,

$$\hat{\beta}_j \sim N\left[\beta_j, \operatorname{var}(\hat{\beta}_j)\right] \qquad j = 1,2,3,\cdots,k \tag{4-8}$$

If we typify

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\operatorname{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N\left[0,1\right] \qquad j = 1,2,3,\cdots,k \tag{4-9}$$

The claim for normality is usually made on the basis of the Central Limit Theorem (*CLT*), but this is restrictive in some cases. That is to say, normality cannot always be assumed. In any application, whether normality of *u* can be assumed is really an empirical matter. It is often the case that using a transformation, i.e. taking logs, yields a distribution that is closer to normality, which is easy to handle from a mathematical point of view. Large samples will allow us to drop normality without affecting the results too much.

Under the *CLM* assumptions 1 through 9, we obtain a Student's *t* distribution

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k} \tag{4-10}$$

where *k* is the number of unknown parameters in the population model (*k*-1 slope parameters and the intercept, β_1). The expression (4-10) is important because it allows us to test a hypothesis on β_j.

If we compare (4-10) with (4-9), we see that the Student's *t* distribution derives from the fact that the parameter σ in $sd(\hat{\beta}_j)$ has been replaced by its estimator $\hat{\sigma}$, which is a random variable. Thus, the degrees of freedom of *t* are *n*-1-*k* corresponding to the degrees of freedom used in the estimation of $\hat{\sigma}^2$.

When the degrees of freedom (*df*) in the *t* distribution are large, the *t* distribution approaches the standard normal distribution. In figure 4.2, the density function for normal and *t* distributions for different *df* are represented. As can be seen, the t density functions are flatter (platycurtic) and the tails are wider than normal density function, but as *df* increases, t density functions are closer to the normal density. In fact, what happens is that the t distribution takes into account that σ^2 is estimated because it is unknown. Given this uncertainty, the t distribution extends more than the normal one. However, as the *df* grows the t-distribution is nearer to the normal distribution because the uncertainty of not knowing σ^2 decreases.

Therefore, the following convergence in distribution should be kept in mind:

$$t_n \xrightarrow[n\to\infty]{} N(0,1) \qquad (4\text{-}11)$$

Thus, when the number of degrees of freedom of a Student's t tends to infinity, the t distribution converges towards a distribution N(0.1). In the context of testing a hypothesis, if the sample size grows, so will the degrees of freedom. This means that for large sizes the normal distribution can be used to test hypothesis with one unique restriction, even when you do not know the population variance. As a practical rule, when the *df* are larger than 120, we can take the critical values from the normal distribution.

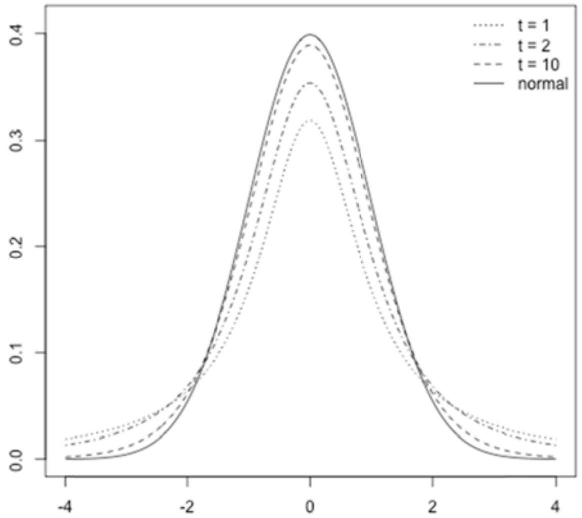


FIGURE 4.2. Density functions: normal and *t* for different degrees of freedom.

Consider the null hypothesis,

$$H_0 : \beta_j = 0$$

Since β_j measures the partial effect of x_j on y after controlling for all other independent variables, $H_0 : \beta_j = 0$ means that, once x_2, x_3, $\ldots$,x_{j-1}, x_{j+1},..., x_k have been accounted for, x_j has *no effect* on y. This is called a *significance test*. The statistic we use to test $H_0 : \beta_j = 0$, against any alternative, is called the *t statistic* or the *t ratio* of $\hat{\beta}_j$ and is expressed as

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

In order to test $H_0 : \beta_j = 0$, it is natural to look at our unbiased estimator of β_j, $\hat{\beta}_j$. In a given sample $\hat{\beta}_j$ will never be exactly zero, but a small value will indicate that the null hypothesis could be true, whereas a large value will indicate a false null hypothesis. The question is: how far is $\hat{\beta}_j$ from zero?

We must recognize that there is a sampling error in our estimate $\hat{\beta}_j$, and thus the size of $\hat{\beta}_j$ must be weighted against its sampling error. This is precisely what we do when we use $t_{\hat{\beta}_j}$, since this statistic measures how many standard errors $\hat{\beta}_j$ is away from zero. In order to determine a rule for rejecting H_0, we need to decide on the relevant *alternative hypothesis*. There are three possibilities: one-tail alternative hypotheses (right and left tail), and two-tail alternative hypothesis.

One-tail alternative hypothesis: right

First, let us consider the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j > 0$$

This is a *positive significance test*. In this case, the decision rule is the following:

Decision rule

$$\begin{array}{lll} \text{If} & t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k} & \text{reject} \quad H_0 \\ \text{If} & t_{\hat{\beta}_j} < t^{\alpha}_{n-k} & \text{not reject } H_0 \end{array} \qquad (4\text{-}12)$$

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j > 0$ at α when $t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$ as can be seen in figure 4.3. It is very clear that to reject H_0 against $H_1 : \beta_j > 0$, we must get a positive $t_{\hat{\beta}_j}$. A negative $t_{\hat{\beta}_j}$, no matter how large, provides no evidence in favor of $H_1 : \beta_j > 0$. On the other hand, in order to obtain t^{α}_{n-k} in the t statistical table, we only need the significance level α and the degrees of freedom.

It is important to remark that as α decreases, t^{α}_{n-k} increases.

To a certain extent, the classical approach is somewhat arbitrary, since we need to choose α in advance, and eventually H_0 is either rejected or not.

In figure 4.4, the alternative approach is represented. As can be seen by observing the figure, the determination of the *p*-value is the inverse operation to find the value of the statistical tables for a given significance level. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, while the null hypothesis is not rejected when α<*p*-value.

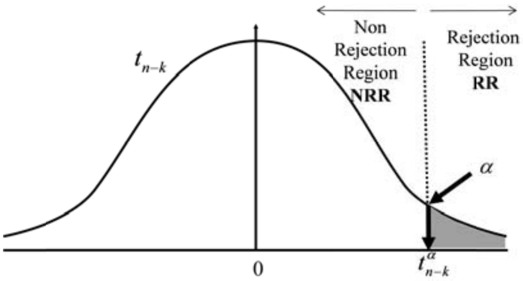
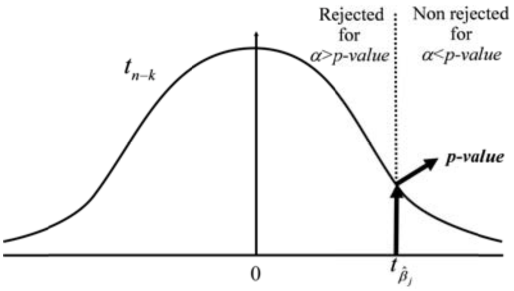


FIGURE 4.3. Rejection region using *t*: right-tail alternative hypothesis.



t_{n-k}
Rejected for α>p-value
Non rejected for α<p-value
p-value
0
$t_{\hat{\beta}_j}$

FIGURE 4.4. *p-value* using *t*: right-tail alternative hypothesis.

EXAMPLE 4.1 Is the marginal propensity to consume smaller than the average propensity to consume?

As seen in example 1.1, testing the 3rd proposition of the Keynesian consumption function in a linear model, is equivalent to testing whether the intercept is significatively greater than 0. That is to say, in the model

$$cons = \beta_1 + \beta_2 inc + u$$

we must test whether

$$\beta_1 > 0$$

With a random sample of 42 observations, the following results have been obtained

$$\widehat{cons}_i = \underset{(0.350)}{0.41} + \underset{(0.062)}{0.843} inc_i$$

The numbers in parentheses, below the estimates, are standard errors *(se*) of the estimators.

The question we pose is the following: is the third proposition of the Keynesian theory admissible? Next, we answer this question.

1) In this case, the null and alternative hypotheses are the following:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

2) The test statistic is:

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{0.41}{0.35} = 1.171$$

3) Decision rule

It is useful to use several significance levels. Let us begin with a significance level of 0.10 because the value of *t* is relatively small (smaller than 1.5). In this case, the degrees of freedom are 40 (42 observations minus 2 estimated parameters). If we look at the *t* statistical table (row 40 and column 0.10, or 0.20, in statistical tables with one tail, or two tails, respectively), we find $t_{40}^{0.10} = 1.303$

As t<1.303, we do not reject H_0 for α=0.10, and therefore we cannot reject for α=0.05 ($t_{40}^{0.05} = 1.684$) or α=0.01 ($t_{40}^{0.01} = 2.423$), as can been in figure 4.5. In this figure, the rejection region corresponds to α=0.10. Therefore, we cannot reject H_0 in favor H_1. In other words, the sample data are not consistent with Keynes's proposition 3.

In the alternative approach, as can be seen in figure 4.6, the *p*-value corresponding to a $t_{\hat{\beta}_1}$ =1.171 for a *t* with 40 *df* is equal to 0.124. For α<0.124 - for example, 0.10, 0.05 and 0.01-, H_0 is not rejected.

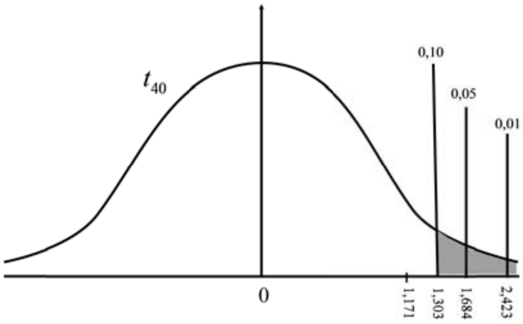


FIGURE 4.5. Example 4.1: Rejection region using *t* with a right-tail alternative hypothesis.

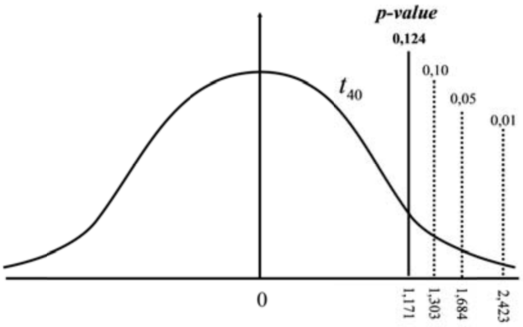


FIGURE 4.6. Example 4.1: *p*-value using *t* with right-tail alternative hypothesis.

One-tail alternative hypothesis: left

Consider now the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j < 0$$

This is a *negative significance test.*

In this case, the decision rule is the following:

Decision rule

$$\begin{array}{llll} \text{If} & t_{\hat{\beta}_j} \leq -t_{n-k}^{\alpha} & \text{reject} & H_0 \\ \text{If} & t_{\hat{\beta}_j} > -t_{n-k}^{\alpha} & \text{not reject} & H_0 \end{array} \qquad (4\text{-}13)$$

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j < 0$ at a given α when $t_{\hat{\beta}_j} \leq -t_n^{\alpha}$, as can be seen in figure 4.7. It is very clear that to reject H_0 against $H_1 : \beta_j < 0$, we must get a negative $t_{\hat{\beta}_j}$. A positive $t_{\hat{\beta}_j}$, no matter how large it is, provides no evidence in favor of $H_1 : \beta_j < 0$.

In figure 4.8 the alternative approach is represented. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, while the null hypothesis is not rejected when α<*p*-value.

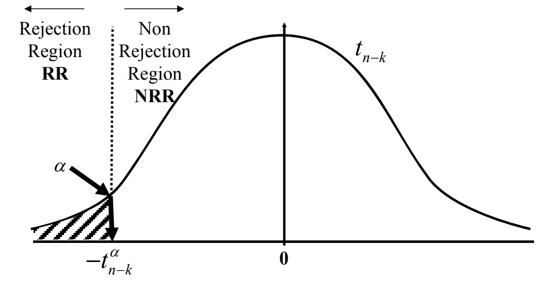


FIGURE 4.7. Rejection region using *t*: left-tail alternative hypothesis.

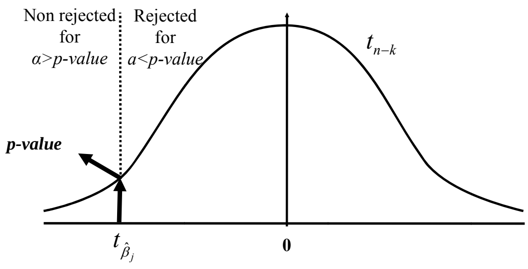


FIGURE 4.8. *p*-value using *t*: left-tail alternative hypothesis.

EXAMPLE 4.2 Has income a negative influence on infant mortality?

The following model has been used to explain the deaths of children under 5 years per 1000 live births (*deathun5*).

$$deathun5 = \beta_1 + \beta_2 gnipc + \beta_3 ilitrate + u$$

where *gnipc* is the gross national income per capita and *ilitrate* is the adult (% 15 and older) illiteracy rate in percentage.

With a sample of 130 countries (workfile *hdr2010*), the following estimation has been obtained:

$$\widehat{deathun5_i} = \underset{(5.93)}{27.91} - \underset{(0.00028)}{0.000826}\, gnipc_i + \underset{(0.183)}{2.043}\, ilitrate_i$$

The numbers in parentheses, below the estimates, are standard errors *(se)* of the estimators.

One of the questions posed by researchers is whether income has a negative influence on infant mortality. To answer this question, the following hypothesis testing is carried out:

The null and alternative hypotheses, and the test statistic, are the following:

$$H_0 : \beta_2 = 0 \qquad H_1 : \beta_2 < 0 \qquad t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{-0.000826}{0.00028} = -2.966$$

Since the *t* value is relatively high, let us start testing with a level of 1%. For α=0.01, $t^{0.01}_{130-1-2} \approx t^{0.01}_{60} = 2.390$. Given that *t*<-2.390, as is shown in figure 4.9, we reject H_0 in favour of H_1. Therefore, the gross national income per capita has an influence that is significantly negative in mortality of children under 5.That is to say, the higher the gross national income per capita the lower the percentage of mortality of children under 5. As H_0 has been rejected for α=0.01, it will also be rejected for levels of 5% and 10%.

In the alternative approach, as can be seen in figure 4.10, the *p*-value corresponding to a $t_{\hat{\beta}_1}$ =-2.966 for a *t* with 61 *df* is equal to 0.0000. For all α>0.0000, such as 0.01, 0.05 and 0.10, H_0 is rejected.

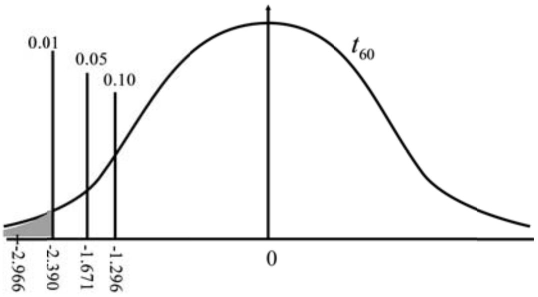


FIGURE 4.9. Example 4.2: Rejection region using *t* with a left-tail alternative hypothesis.

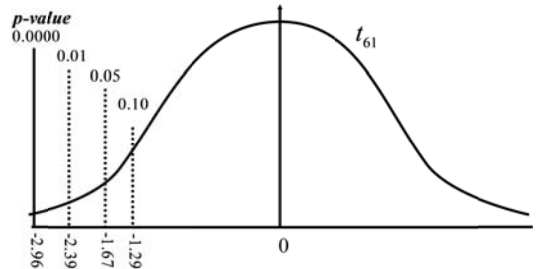


FIGURE 4.10. Example 4.2: *p*-value using *t* with a left-tail alternative hypothesis.

Two-tail alternative hypothesis

Consider now the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j \neq 0$$

This is the relevant alternative when the sign of β_j is not well determined by theory or common sense. When the alternative is two-sided, we are interested in the *absolute value* of the *t* statistic. This is a *significance test.*

In this case, the decision rule is the following:

Decision rule

$$\text{If} \quad \left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2} \quad \text{reject} \quad H_0$$
$$\text{If} \quad \left| t_{\hat{\beta}_j} \right| < t_{n-k}^{\alpha/2} \quad \text{not reject} \quad H_0 \qquad \text{(4-14)}$$

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j < 0$ at α when $\left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2}$, as can be seen in figure 4.11. In this case, in order to reject H_0 against $H_1 : \beta_j \neq 0$, we must obtain a large enough $t_{\hat{\beta}_j}$ which is either positive or negative.

It is important to remark that as α decreases, $t_{n-k}^{\alpha/2}$ increases in absolute value.

In the alternative approach, once the *p*-value has been determined, we know that while H_0 is rejected for any level of significance of α>*p*-value, the null hypothesis is not rejected when α<*p*-value. In this case, the *p*-value is distributed between both tails in a symmetrical way, as is shown in figure 4.12.

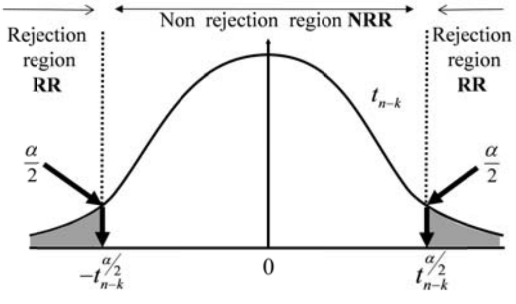


FIGURE 4.11. Rejection region using *t*: two-tail alternative hypothesis.

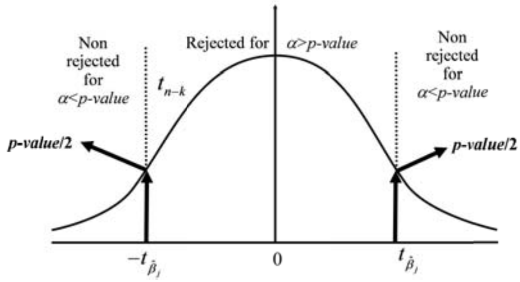


FIGURE 4.12. *p*-value using *t*: two-tail alternative hypothesis.

When a specific alternative hypothesis is not stated, it is usually considered to be two-sided hypothesis testing. If H_0 is rejected in favor of H_1 at a given α, we usually say that "x_j is *statistically significant* at the level α".

EXAMPLE 4.3 Has the rate of crime play a role in the price of houses in an area?

To explain housing prices in an American town, the following model is estimated:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

where *rooms* is the number of rooms of the house, *lowstat* is the percentage of people of "lower status" in the area and *crime* is crimes committed per capita in the area.

The output for the fitted model, using the file *hprice2* (first 55 observations), appears in table 4.2 and has been taken from E-views. The meaning of the first three columns is clear: "t-Statistic" is the outcome to perform a significance test, that is to say, it is the ratio between the "Coefficient" and the "Std error"; and "Prob" is the *p*-value to perform a two-tailed test.

In relation to this model, the researcher questions whether the rate of crime in an area plays a role in the price of houses in that area.

To answer this question, the following procedure has been carried out.

In this case, the null and alternative hypothesis and the test statistic are the following:

$$H_0 : \beta_4 = 0 \qquad t = \frac{\hat{\beta}_4}{se(\hat{\beta}_4)} = \frac{-3854}{960} = -4.016$$
$$H_1 : \beta_4 \neq 0$$

TABLE 4.2. Standard output in the regression explaining house price. *n*=55.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	-15693.61	8021.989	-1.956324	0.0559
ROOMS	6788.401	1210.720	5.606910	0.0000
LOWSTAT	-268.1636	80.70678	-3.322690	0.0017
CRIME	-3853.564	959.5618	-4.015962	0.0002

Since the *t* value is relatively high, let us by start testing with a level of 1%. For α=0.01, $t_{51}^{0.01/2} \approx t_{50}^{0.01/2} = 2.69$. (In the usual statistical tables for *t* distribution, there is no information for each *df* above 20). Given that $|t| > 2.69$, we reject H_0 in favour of H_1. Therefore, crime has a significant influence on housing prices for a significance level of 1% and, thus, of 5% and 10%.

In the alternative approach, we can perform the test with more precision. In table 4.2 we see that the *p*-value for the coefficient of crime is 0.0002. That means that the probability of the *t* statistic being greater than 4.016 is 0.0001 and the probability of *t* being smaller than -4.016 is 0.0001. That is to say, the *p*-value, as shown in Figure 4.13, is distributed in the two tails. As can be seen in this figure, H_0 is rejected for all significance levels greater than 0.0002, such as 0.01, 0.05 and 0.10.

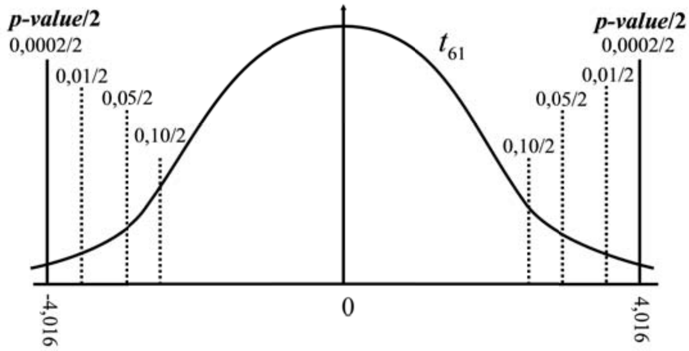


FIGURE 4.13. Example 4.3: *p*-value using *t* with a two-tail alternative hypothesis.

So far we have seen significant tests of one-tail and two-tails, in which a parameter takes the value 0 in H_0. Now we are going to look at a more general case where the parameter in H_0 takes any value:

$$H_0 : \beta_j = \beta_j^0$$

Thus, the appropriate *t* statistic is

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)}$$

As before, $t_{\hat{\beta}_j}$ measures how many estimated standard deviations $\hat{\beta}_j$ is away from the hypothesized value of β_j^0.

EXAMPLE 4.4 Is the elasticity expenditure in fruit/income equal to 1? Is fruit a luxury good?

To answer these questions, we are going to use the following model for the expenditure in *fruit*:

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 househsize + \beta_4 punders + u$$

where *inc* is disposable income of household, *househsize* is the number of household members and *punder5* is the proportion of children under five in the household.

As the variables *fruit* and *inc* appear expressed in natural logarithms, then β_2 is the expenditure in fruit/income elasticity. Using a sample of 40 households (workfile *demand*), the results of table 4.3 have been obtained.

TABLE 4.3. Standard output in a regression explaining expenditure in fruit. *n*=40.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	-9.767654	3.701469	-2.638859	0.0122
LN(INC)	2.004539	0.512370	3.912286	0.0004
HOUSEHSIZE	-1.205348	0.178646	-6.747147	0.0000
PUNDER5	-0.017946	0.013022	-1.378128	0.1767

Is the expenditure in fruit/income elasticity equal to 1?

To answer this question, the following procedure has been carried out:

In this case, the null and alternative hypothesis and the test statistic are the following:

$$H_0: \beta_2 = 1 \qquad H_1: \beta_2 \neq 1 \qquad t = \frac{\hat{\beta}_2 - \beta_2^0}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \frac{2.005 - 1}{0.512} = 1.961$$

For α=0.10, we find that $t_{36}^{0.10/2} \approx t_{35}^{0.10/2} = 1.69$. As $|t|$>1.69, we reject H_0. For α=0.05, $t_{36}^{0.05/2} \approx t_{35}^{0.05/2} = 2.03$. As $|t|$<2.03, we do not reject H_0 for α=0.05, nor for α=0.01. Therefore, we reject that the expenditure on fruit/income elasticity is equal to 1 for α=0.10, but we cannot reject it for α=0.05, nor for α=0.01.

Is fruit a luxury good?

According to economic theory, a commodity is a luxury good when its expenditure elasticity with respect to income is higher than 1. Therefore, to answer to the second question, and taking into account that the t statistic is the same, the following procedure has been carried out:

$$H_0: \beta_2 = 1 \qquad H_1: \beta_2 > 1.$$

For α=0.10, we find that $t_{36}^{0.10} \approx t_{35}^{0.10} = 1.31$. As t>1.31, we reject H_0 in favour of H_1. For α=0.05, $t_{36}^{0.05} \approx t_{35}^{0.05} = 1.69$. As t>1.69, we reject H_0 in favour of H_1. For α=0.01, $t_{36}^{0.01} \approx t_{35}^{0.01} = 2.44$. As t<2.44, we do not reject H_0. Therefore, fruit is a luxury good for α=0.10 and α=0.05, but we cannot reject H_0 in favour of H_1 for α=0.01.

EXAMPLE 4.5 Is the Madrid stock exchange market efficient?

Before answering this question, we will examine some previous concepts. The *rate of return of an asset* over a period of time is defined as the percentage change in the value invested in the asset during that period of time. Let us now consider a specific asset: a share of an industrial company acquired in a Spanish stock market at the end of one year and remains until the end of next year. Those two moments of time will be denoted by *t*-1 and *t* respectively. The rate of return of this action within that year can be expressed by the following relationship:

$$RA_t = \frac{\Delta P_t + D_t + A_t}{P_{t-1}} \tag{4-15}$$

where P_t: is the share price at the end of period *t*, D_t: are the dividends received by the share during the period *t*, and A_t: is the value of the rights that eventually corresponded to the share during the period *t*

Thus, the numerator of (4-15) summarizes the three types of capital gains that have been received for the maintenance of a share in year *t*; that is to say, an increase or decrease in quotation, dividends and rights on capital increase. Dividing by P_{t-1}, we obtain the rate of profit on share value at the end of the previous period. Of these three components, the most important one is the increase in quotation. Considering only that component, the yield rate of the action can be expressed by

$$RA1_t = \frac{\Delta P_t}{P_{t-1}} \tag{4-16}$$

or, alternatively if we use a relative rate of variation, by

$$RA2_t = \Delta \ln P_t \tag{4-17}$$

In the same way as Ra_t represents the rate of return of a particular share in either of the two expressions, we can also calculate the rate of return of all shares listed in the stock exchange. The latter rate of return, which will be denoted by RM_t, is called the market rate of return.

So far we have considered the rate of return in a year, but we can also apply expressions such as (4-16), or (4-17), to obtain daily rates of return. It is interesting to know whether the rates of return in the past are useful for predicting rates of return in the future. This question is related to the concept of market efficiency. A market is *efficient* if prices incorporate all available information, so there is no possibility of making abnormal profits by using this information.

In order to test the efficiency of a market, we define the following model, using daily rates of return defined by (4-16):

$$rmad92_t = \beta_1 + \beta_2 rmad92_{t-1} + u_t \tag{4-18}$$

If a market is efficient, then the parameter β_2 of the previous model must be 0. Let us now compare whether the Madrid Stock Exchange is efficient as a whole.

The model (4-18) has been estimated with daily data from the Madrid Stock Exchange for 1992, using file *bolmadef*. The results obtained are the following:

$$\widehat{rmad92_t} = \underset{(0.0007)}{-0.0004} + \underset{(0.0629)}{0.1267}\, rmad92_{t-1}$$

$$R^2=0.0163 \qquad n=247$$

The results are paradoxical. On the one hand, the coefficient of determination is very low (0.0163), which means that only 1.63% of the total variance of the rate of return is explained by the previous day's rate of return. On the other hand, the coefficient corresponding to the rate of significance of the previous day is statistically significant at a level of 5% but not at a level of 1% given that the t statistic is equal to 0.1267/0.0629=2.02, which is slightly larger in absolute value than $t_{245}^{0.01} \simeq t_{60}^{0.01}$=2.00. The reason for this apparent paradox is that the sample size is very high. Thus, although the impact of the explanatory variable on the endogenous variable is relatively small (as indicated by the coefficient of determination), this finding is significant (as evidenced by the statistical t) because the sample is sufficiently large.

To answer the question as to whether the Madrid Stock Exchange is an efficient market, we can say that it is not entirely efficient. However, this response should be qualified. In financial economics there is a dependency relationship of the rate of return of one day with respect to the rate corresponding to the previous day. This relationship is not very strong, although it is statistically significant in many world stock markets due to market frictions. In any case, market players cannot exploit this phenomenon, and thus the market is not inefficient, according to the above definition of the concept of efficiency.

EXAMPLE 4.6 Is the rate of return of the Madrid Stock Exchange affected by the rate of return of the Tokyo Stock Exchange?

The study of the relationship between different stock markets (NYSE, Tokyo Stock Exchange Madrid Stock Exchange, London Stock Exchange, etc.) has received much attention in recent years due to the greater freedom in the movement of capital and the use of foreign markets to reduce the risk in portfolio management. This is because the absence of perfect market integration allows diversification of risk. In any case, there is a world trend toward a greater global integration of financial markets in general and stock markets in particular.

If markets are efficient, and we have seen in example 4.5 that they are, the *innovations* (new information) will be reflected in the different markets for a period of 24 hours.

It is important to distinguish between two types of innovations: a) *global innovations*, which is news generated around the world and has an influence on stock prices in all markets, b) *specific innovations*, which is the information generated during a 24 hour period and only affects the price of a particular market. Thus, information on the evolution of oil prices can be considered as a global innovation, while a new financial sector regulation in a country would be considered a specific innovation.

According to the above discussion, stock prices quoted at a session of a particular stock market are affected by the global innovations of a different market which had closed earlier. Thus, global innovations included in the Tokyo market will influence the market prices of Madrid on the same day.

The following model shows the transmission of effects between the Tokyo Stock Exchange and the Madrid Stock Exchange in 1992:

$$rmad92_t = \beta_1 + \beta_2 rtok92_t + u_t \tag{4-19}$$

where $rmad92_t$ is the rate of return of the Madrid Stock Exchange in period t and $rtok92_t$ is the rate of return of the Tokyo Stock Exchange in period t. The rates of return have been calculated according to (4-16).

In the working file *madtok* you can find general indices of the Madrid Stock Exchange and the Tokyo Stock Exchange during the days both exchanges were open simultaneously in 1992. That is, we eliminated observations for those days when any one of the two stock exchanges was closed. In total, the number of observations is 234, compared to the 247 and 246 days that the Madrid and Tokyo Stock Exchanges were open.

The estimation of the model (4-19) is as follows:

$$\widehat{rmad92_t} = -\underset{(0.0007)}{0.0005} + \underset{(0.0375)}{0.1244} rtok92_t$$

$$R^2=0.0452 \qquad n=235$$

Note that the coefficient of determination is relatively low. However, for testing H_0: $\beta_2=0$, the statistic t = (0.1244/0.0375) = 3.32, which implies that we reject the hypothesis that the rate of return of the Tokyo Stock Exchange has no effect on the rate of return of the Madrid Stock Exchange, for a significance level of 0.01.

Once again we find the same apparent paradox which appeared when we analyzed the efficiency of the Madrid Stock Exchange in example 4.5 except for one difference. In the latter case, the rate of return from the previous day appeared as significant due to problems arising in the elaboration of the general index of the Madrid Stock Exchange.

Consequently, the fact that the null hypothesis is rejected implies that there is empirical evidence supporting the theory that global innovations from the Tokyo Stock Exchange are transmitted to the quotes of the Madrid Stock Exchange that day.

4.2.2 Confidence intervals

Under the *CLM*, we can easily construct a *confidence interval* (*CI*) for the population parameter, β_j. *CI* are also called interval estimates because they provide a range of likely values for β_j, and not just a point estimate.

The *CI* is built in such a way that the unknown parameter is contained within the range of the *CI* with a previously specified probability.

By using the fact that

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k}$$

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \le t_{n-k}^{\alpha/2}\right] = 1-\alpha$$

Operating to put the unknown β_j alone in the middle of the interval, we have

$$\Pr\left[\hat{\beta}_j - se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2} \le \beta_j \le \hat{\beta}_j + se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha$$

Therefore, the lower and upper bounds of a (1-α) *CI* respectively are given by

$$\underline{\beta}_j = \hat{\beta}_j - se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}$$

$$\overline{\beta}_j = \hat{\beta}_j + se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}$$

If random samples were obtained over and over again with $\underline{\beta}_j$, and $\overline{\beta}_j$ computed each time, then the (unknown) population value would lie in the interval ($\underline{\beta}_j$, $\overline{\beta}_j$) for (1 − α)% of the samples. Unfortunately, for the single sample that we use to construct *CI*, we do not know whether β_j is actually contained in the interval.

Once a *CI* is constructed, it is easy to carry out two-tailed hypothesis tests. If the null hypothesis is $H_0 : \beta_j = a_j$, then H_0 is rejected against $H_1 : \beta_j \neq a_j$ at (say) the 5% significance level if, and only if, a_j is *not* in the 95% *CI.*

To illustrate this matter, in figure 4.14 we constructed confidence intervals of 90%, 95% and 99%, for the marginal propensity to consumption -β_2- corresponding to example 4.1.

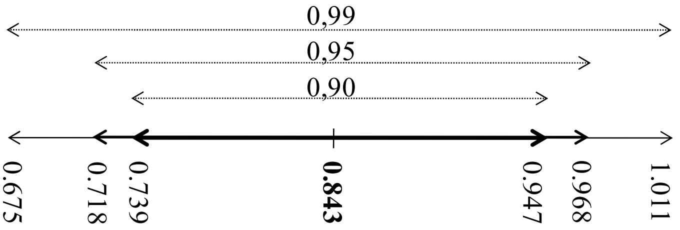


FIGURE 4.14. Confidence intervals for marginal propensity to consume in example 4.1.

4.2.3 Testing hypotheses about a single linear combination of the parameters

In many applications we are interested in testing a hypothesis involving more than one of the population parameters. We can also use the *t* statistic to test a single linear combination of the parameters, where two or more parameters are involved.

There are two different procedures to perform the test with a single linear combination of parameters. In the first, the standard error of the linear combination of parameters corresponding to the null hypothesis is calculated using information on the covariance matrix of the estimators. In the second, the model is reparameterized by introducing a new parameter derived from the null hypothesis and the reparameterized model is then estimated; testing for the new parameter indicates whether the null hypothesis is rejected or not. The following example illustrates both procedures.

EXAMPLE 4.7 Are there constant returns to scale in the chemical industry?

To examine whether there are constant returns to scale in the chemical sector, we are going to use the Cobb-Douglas production function, given by

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u \qquad (4\text{-}20)$$

In the above model parameters β_2 and β_3 are elasticities (output/labor and output/capital).

Before making inferences, remember that *returns to scale* refers to a technical property of the production function examining changes in output subsequent to a change of the same proportion in all inputs, which are labor and capital in this case. If output increases by that same proportional change then there are *constant returns to scale*. Constant returns to scale imply that if the factors *labor* and *capital* increase at a certain rate (say 10%), output will increase at the same rate (e.g., 10%). If output increases by more than that proportion, there are *increasing returns to scale*. If output increases by less than that proportional change, there are *decreasing returns to scale*. In the above model, the following occurs

- if β_2+β_3=1, there are *constant returns to scale.*

- if β_2+β_3>1, there are *increasing returns to scale.*

- if β_2+β_3<1, there are *decreasing returns to scale.*

Data used for this example are a sample of 27 companies of the primary metal sector (workfile *prodmet*), where *output* is gross value added, *labor* is a measure of labor input, and *capital* is the gross value of plant and equipment. Further details on construction of the data are given in Aigner, *et al.* (1977) and in Hildebrand and Liu (1957); these data were used by Greene in 1991. The results obtained in the estimation of model (4-20), using any econometric software available, appear in table 4.4.

TABLE 4.4. Standard output of the estimation of the production function: model (4-20).

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	0.602999	0.125954	4.787457	0.0001
ln(*capital*)	0.375710	0.085346	4.402204	0.0002

To answer the question posed in this example, we must test

$$H_0 : \beta_2 + \beta_3 = 1$$

against the following alternative hypothesis

$$H_1 : \beta_2 + \beta_3 \neq 1$$

According to H_0, it is stated that $\beta_2 + \beta_3 - 1 = 0$. Therefore, the *t* statistic must now be based on whether the estimated sum $\hat{\beta}_2 + \hat{\beta}_3 - 1$ is sufficiently different from 0 to reject H_0 in favor of H_1.

Two procedures will be used to test this hypothesis. In the first, the covariance matrix of the estimators is used. In the second, the model is reparameterized by introducing a new parameter.

Procedure: using covariance matrix of estimators

According to H_0, it is stated that $\beta_2 + \beta_3 - 1 = 0$. Therefore, the *t* statistic must now be based on whether the estimated sum $\hat{\beta}_2 + \hat{\beta}_3 - 1$ is sufficiently different from 0 to reject H_0 in favor of H_1. To account for the sampling error in our estimators, we standardize this sum by dividing by its standard error:

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$

Therefore, if $t_{\hat{\beta}_2+\hat{\beta}_3}$ is large enough, we will conclude, in a two side alternative test, that there are not *constant returns to scale*. On the other hand, if $t_{\hat{\beta}_2+\hat{\beta}_3}$ is positive and large enough, we will reject, in a one side alternative test (right), H_0 in favour of $H_1 : \beta_2 + \beta_3 > 1$. Therefore, there are *increasing returns to scale*.

On the other hand , we have

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\widehat{\text{var}(\hat{\beta}_2 + \hat{\beta}_3)}}$$

where

$$\widehat{\text{var}(\hat{\beta}_2 + \hat{\beta}_3)} = \widehat{\text{var}(\hat{\beta}_2)} + \widehat{\text{var}(\hat{\beta}_3)} + 2 \times \widehat{\text{covar}(\hat{\beta}_2, \hat{\beta}_3)}$$

Hence, to compute $se(\hat{\beta}_2 + \hat{\beta}_3)$ you need information on the estimated covariance of estimators. Many econometric software packages (such as e-views) have an option to display estimates of the covariance matrix of the estimator vector '. In this case, the covariance matrix obtained appears in table 4.5. Using this information, we have

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{0.015864 + 0.007284 - 2 \times 0.009616} = 0.0626$$

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{se(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{-0.02129}{0.0626} = -0.3402$$

TABLE 4.5. Covariance matrix in the production function.

	constant	ln(*labor*)	ln(*capital*)
constant	0.106786	-0.019835	0.001189
ln(*labor*))	-0.019835	0.015864	-0.009616
ln(*capital*)	0.001189	-0.009616	0.007284

Given that *t*=0.3402, it is clear that we cannot reject the existence of constant returns to scale for the usual significance levels. Given that the t statistic is negative, it makes no sense to test whether there are increasing returns to scale

Procedure: reparameterizing the model by introducing a new parameter

It is easier to perform the test if we apply the second procedure. A different model is estimated in this procedure, which directly provides the standard error of interest. Thus, let us define:

$$\theta = \beta_2 + \beta_3 - 1$$

thus, the null hypothesis that there are *constant returns to scale* is equivalent to saying that $H_0 : \theta = 0$.

From the definition of θ, we have $\beta_2 = \theta - \beta_3 + 1$. Substituting β_2 in the original equation:

$$\ln(output) = \beta_1 + (\theta - \beta_3 + 1)\ln(labor) + \beta_3 \ln(capital) + u$$

Hence,

$$\ln(output / labor) = \beta_1 + \theta \ln(labor) + \beta_3 \ln(capital / labor) + u$$

Therefore, to test whether there are constant returns to scale is equivalent to carrying out a significance test on the coefficient of ln(*labor*) in the previous model. The strategy of rewriting the model so that it contains the parameter of interest works in all cases and is usually easy to implement. If we apply this transformation to this example, we obtain the results of Table 4.6.

As can be seen we obtain the same result:

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{se(\hat{\theta})} = -0.3402$$

TABLE 4.6. Estimation output for the production function: reparameterized model.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	-0.021290	0.062577	-0.340227	0.7366
ln(*capital/labor*)	0.375710	0.085346	4.402204	0.0002

EXAMPLE 4.8 Advertising or incentives?

The *Bush Company* is engaged in the sale and distribution of gifts imported from the Near East. The most popular item in the catalog is the *Guantanamo* bracelet, which has some relaxing properties. The sales agents receive a commission of 30% of total sales amount. In order to increase sales without expanding the sales network, the company established special incentives for those agents who exceeded a sales target during the last year.

Advertising spots were radio broadcasted in different areas to strengthen the promotion of sales. In those spots special emphasis was placed on highlighting the well-being of wearing a *Guantanamo* bracelet.

The manager of the *Bush Company* wonders whether a dollar spent on special incentives has a higher incidence on sales than a dollar spent on advertising. To answer that question, the company's econometrician suggests the following model to explain *sales*:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u$$

where *incent* are incentives to the salesmen and *advert* are expenditures in advertising. The variables *sales*, *incent* and *advert* are expressed in thousands of dollars.

Using a sample of 18 sale areas (workfile *advincen*), we have obtained the output and the covariance matrix of the coefficients that appear in table 4.7 and in table 4.8 respectively.

TABLE 4.7. Standard output of the regression for example 4.8.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	396.5945	3548.111	0.111776	0.9125
advert	18.63673	8.924339	2.088304	0.0542
incent	30.69686	3.604420	8.516448	0.0000

TABLE 4.8. Covariance matrix for example 4.8.

	C	*ADVERT*	*INCENT*
constant	12589095	-26674	-7101
advert	-26674	79.644	2.941
incent	-7101	2.941	12.992

In this model, the coefficient β_2 indicates the increase in sales produced by a dollar increase in spending on advertising, while β_3 indicates the increase caused by a dollar increase in the special incentives, holding fixed in both cases the other regressor.

To answer the question posed in this example, the null and the alternative hypothesis are the following:

$$H_0 : \beta_3 - \beta_2 = 0$$
$$H_1 : \beta_3 - \beta_2 > 0$$

The t statistic is built using information about the covariance matrix of the estimators:

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{se(\hat{\beta}_3 - \hat{\beta}_2)}$$

$$se(\hat{\beta}_3 - \hat{\beta}_2) = \sqrt{79.644 + 12.992 - 2 \times 2.941} = 9.3142$$

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{se(\hat{\beta}_3 - \hat{\beta}_2)} = \frac{30.697 - 18.637}{9.3142} = 1.295$$

For α=0.10, we find that $t_{15}^{0.10} = 1.341$. As t<1.341, we do not reject H_0 for α=0.10, nor for α=0.05 or α=0.01. Therefore, there is no empirical evidence that a dollar spent on special incentives has a higher incidence on sales than a dollar spent on advertising.

EXAMPLE 4.9 Testing the hypothesis of homogeneity in the demand for fish

In the case study in chapter 2, models for demand for dairy products have been estimated from cross-sectional data, using disposable income as an explanatory variable. However, the price of the product itself and, to a greater or lesser extent, the prices of other goods are determinants of the demand. The demand analysis based on cross sectional data has precisely the limitation that it is not possible to examine the effect of prices on demand because prices remain constant, since the data refer to the same point in time. To analyze the effect of prices it is necessary to use time series data or, alternatively, panel data. We will briefly examine some aspects of the theory of demand for a good and then move to the estimation of a demand function with time series data. As a postscript to this case, we will test one of the hypotheses which, under certain circumstances, a theoretical model must satisfy.

The demand for a commodity - say good j - can be expressed, according to an optimization process carried out by the consumer, in terms of disposable income, the price of the good and the prices of the other goods. Analytically:

$$q_j = f_j(p_1, p_2, \cdots, p_j, \cdots, p_m, di) \qquad (4\text{-}21)$$

where

- *di* is the disposable income of the consumer.
- $p_1, p_2, \cdots, p_j, \cdots p_m$ are the prices of the goods which are taken into account by consumers when they acquire the good j.

Logarithmic models are attractive in studies on demand,, because the coefficients are directly elasticities. The log model is given by

$$\ln(q_j) = \beta_1 + \beta_2 \ln(p_1) + \beta_3 \ln(p_2) + \cdots + \beta_j \ln(p_j) + \cdots + \beta_{m+1} \ln(p_m) + \beta_{m+2} \ln(R) + u \qquad (4\text{-}22)$$

It is clear to see that all β coefficients*,* excluding the constant term, are elasticities of different types and therefore are independent of the units of measurement for the variables. When there is no money illusion, if all prices and income grow at the same rate, the demand for a good is not affected by these changes. Thus, assuming that prices and income are multiplied by λ, if the consumer has no money illusion, the following should be satisfied

$$f_j(\lambda p_1,\lambda p_2,\cdots,\lambda p_j,\cdots,\lambda p_m,\lambda R) = f_j(p_1,p_2,\cdots,p_j,\cdots p_m,di) \qquad (4\text{-}23)$$

From a mathematical point of view, the above condition implies that the demand function must be homogeneous of degree 0. This condition is called the *restriction of homogeneity*. Applying Euler's theorem, the restriction of homogeneity in turn implies that the sum of the demand/income elasticity and of all demand/price elasticities is zero, i.e.:

$$\sum_{h=1}^{m}\varepsilon_{q_j/p_h} + \varepsilon_{q_j/R} = 0 \qquad (4\text{-}24)$$

This restriction applied to the logarithmic model (4-22) implies that

$$\beta_2 + \beta_3 + \cdots + \beta_j + \cdots + \beta_{m+1} + \beta_{m+2} = 0 \qquad (4\text{-}25)$$

In practice, when estimating a demand function, the prices of many goods are not included, but only those that are closely related, either because they are complementary or substitute goods. It is also well known that the budgetary allocation of spending is carried out in several stages.

Next, the demand for fish in Spain will be studied by using a model similar to (4-22). Let us consider that in a first assignment, the consumer distributes its income between total consumption and savings. In a second stage, the consumption expenditure by function is performed taking into account the total consumption and the relevant prices in each function. Specifically, we assume that the only relevant price in the demand for fish is the price of the good (fish) and the price of the most important substitute (meat).

Given the above considerations, the following model is formulated:

$$\ln(fish) = \beta_1 + \beta_2\ln(fishpr) + \beta_3\ln(meatpr) + \beta_4\ln(cons) + u \qquad (4\text{-}26)$$

where *fish* is fish expenditure at constant prices, *fishpr* is the price of fish, *meatpr* is the price of meat and *cons* is total consumption at constant prices.

The workfile *fishdem* contains information about this series for the period 1964-1991. Prices are index numbers with 1986 as a base, and *fish* and *cons* are magnitudes at constant prices with 1986 as a base also. The results of estimating model (4-26) are as follows:

$$\widehat{\ln(fish)} = \underset{(2.30)}{7.788} - \underset{(0.133)}{0.460}\ln(fishpr) + \underset{(0.112)}{0.554}\ln(meatpr) + \underset{(0.137)}{0.322}\ln(cons)$$

As can be seen, the signs of the elasticities are correct: the elasticity of demand is negative with respect to the price of the good, while the elasticities with respect to the price of the substitute good and total consumption are positive

In model (4-26) the homogeneity restriction implies the following null hypothesis:

$$\beta_2 + \beta_3 + \beta_4 = 0 \qquad (4\text{-}27)$$

To carry out this test, we will use a similar procedure to the one used in example 4.6. Now, the parameter θ is defined as follows

$$\theta = \beta_2 + \beta_3 + \beta_4 \qquad (4\text{-}28)$$

Setting $\beta_2 = \theta - \beta_3 - \beta_4$, the following model has been estimated:

$$\ln(fish) = \beta_1 + \theta\ln(fishpr) + \beta_3\ln(meatpr/fishpr) + \beta_4\ln(cons/fishpr) + u \qquad (4\text{-}29)$$

The results obtained were the following:

$$\widehat{\ln(fish_i)} = \underset{(2.30)}{7.788} - \underset{(0.1334)}{0.4596}\ln(fishpr_i) + \underset{(0.112)}{0.554}\ln(meatpr_i) + \underset{(0.137)}{0.322}\ln(cons_i)$$

Using (4-28), testing the null hypothesis (4-27) is equivalent to testing that the coefficient of ln(*fishpr*) in (4-29) is equal to 0. Since the t statistic for this coefficient is equal to -3.44 and $t_{24}^{0.01/2}$ =2.8, we reject the hypothesis of homogeneity regarding the demand for fish.

4.2.4 Economic importance versus statistical significance

Up until now we have emphasized statistical significance. However, it is important to remember that we should pay attention to the magnitude and the sign of the estimated coefficient in addition to *t* statistics.

Statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$, whereas the economic significance of a variable is related to the size (and sign) of $\hat{\beta}_j$. Too much focus on statistical significance can lead to the false conclusion that a variable is "important" for explaining *y*, even though its estimated effect is modest.

Therefore, even if a variable is statistically significant, you need to discuss the magnitude of the estimated coefficient to get an idea of its practical or economic importance.

4.3 Testing multiple linear restrictions using the *F* test.

So far, we have only considered hypotheses involving a single restriction. But frequently, we wish to test multiple hypotheses about the underlying parameters $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$.

In multiple linear restrictions, we will distinguish three types: *exclusion restrictions*, *model significance* and *other linear restrictions*.

4.3.1 Exclusion restrictions

Null and alternative hypotheses; unrestricted and restricted model

We begin with the leading case of testing whether a set of independent variables has no partial effect on the dependent variable, *y*. These are called *exclusion restrictions*. Thus, considering the model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u \qquad (4\text{-}30)$$

the null hypothesis in a typical example of exclusion restrictions could be the following:

$$H_0 : \beta_4 = \beta_5 = 0$$

This is an example of a set of *multiple restrictions*, because we are putting more than one restriction on the parameters in the above equation. A test of multiple restrictions is called a *joint* hypothesis test.

The alternative hypothesis can be expressed in the following way

$$H_1\text{: } H_0 \text{ is not true}$$

It is important to remark that we test the above H_0 jointly, not individually. Now, we are going to distinguish between *unrestricted* (*UR*) and *restricted* (*R*) models. The unrestricted model is the reference model or initial model. In this example the unrestricted model is the model given in (4-30). The restricted model is obtained by imposing H_0 on the original model. In the above example, the restricted model is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

By definition, the restricted model always has fewer parameters than the unrestricted one. Moreover, it is always true that

$$RSS_R \geq RSS_{UR}$$

where RSS_R is the *RSS* of the restricted model, and RSS_{UR} is the *RSS* of the unrestricted model. Remember that, because *OLS* estimates are chosen to minimize the sum of squared residuals, the *RSS* never decreases (and generally increases) when certain restrictions (such as dropping variables) are introduced into the model.

The increase in the *RSS* when the restrictions are imposed can tell us something about the likely truth of H_0. If we obtain a large increase, this is evidence against H_0, and this hypothesis will be rejected. If the increase is small, this is not evidence against H_0, and this hypothesis will not be rejected. The question is therefore whether the observed increase in the *RSS* when the restrictions are imposed is large enough, relative to the *RSS* in the unrestricted model, to warrant rejecting H_0.

The answer depends on α, but we cannot carry out the test at a chosen α until we have a statistic whose distribution is known, and is tabulated, under H_0. Thus, we need a way to combine the information in RSS_R and RSS_{UR} to obtain a test statistic with a known distribution under H_0.

Now, let us look at the general case, where the *unrestricted model* is

$$y=\beta_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_kx_k+u \tag{4-31}$$

Let us suppose that there are q exclusion restrictions to test. H_0 states that q of the variables have zero coefficients. Assuming that they are the last q variables, H_0 is stated as

$$H_0:\beta_{k-q+1}=\beta_{k-q+2}=\cdots=\beta_k=0 \tag{4-32}$$

The restricted model is obtained by imposing the q restrictions of H_0 on the unrestricted model.

$$y=\beta_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_{k-q}x_{k-q}+u \tag{4-33}$$

H_1 is stated as

$$H_1\text{: } H_0 \text{ is not true} \tag{4-34}$$

Test statistic: F ratio

The *F* statistic, or *F* ratio, is defined by

$$F=\frac{(RSS_R-RSS_{UR})/q}{RSS_{UR}/(n-k)} \tag{4-35}$$

where RSS_R is the *RSS* of the restricted model, and RSS_{UR} is the *RSS* of the unrestricted model and q is the number of restrictions; that is to say, the number of equalities in the null hypothesis.

In order to use the *F* statistic for a hypothesis testing, we have to know its sampling distribution under H_0 in order to choose the value c for a given α, and determine the rejection rule. It can be shown that, under H_0, and assuming the *CLM* assumptions hold, the *F* statistic is distributed as a Snedecor's *F* random variable with (q,n-k) *df*. We write this result as

$$F\mid H_0\sim F_{q,n-k} \tag{4-36}$$

A Snedecor's F with *q degrees of freedom* in the numerator and *n-k* de *degrees of freedom* in the denominator is equal to

$$F_{q,n-k} = \frac{\chi_q^2 / q}{\chi_{n-k}^2 / n-k} \tag{4-37}$$

where x_q^2 and x_{n-k}^2 are Chi-square distributions that are independent of each other.

In (4-35) we see that the *degrees of freedom* corresponding to RSS_{UR} (df_{UR})are *n-k.* Remember that

$$\hat{\sigma}_{UR}^2 = \frac{RSS_{UR}}{n-k} \tag{4-38}$$

On the other hand, the *degrees of freedom* corresponding to RSS_R (df_R) are *n-k+q,* because in the restricted model *k-q* parameters are estimated. The *degrees of freedom* corresponding to RSS_R-RSS_{UR} are

$(n-k+q)-(n-k)=q$ = *numerator degrees of freedom*=df_R-df_{UR}

Thus, in the numerator of F, the difference in *RSS*′s is divided by q, which is the number of restrictions imposed when moving from the unrestricted to the restricted model. In the denominator of F, RSS_{UR} is divided by df_{UR}. In fact, the denominator of F is simply the unbiased estimator of σ^2 in the unrestricted model.

The F ratio must be greater than or equal to 0, since $SSR_R - SSR_{UR} \geq 0$.

It is often useful to have a form of the F statistic that can be computed from the R^2 of the restricted and unrestricted models.

Using the fact that $RSS_R = TSS(1-R_R^2)$ and $RSS_{UR} = TSS(1-R_{UR}^2)$, we can write (4-35) as the following

$$F = \frac{(R_{UR}^2 - R_R^2)/q}{(1-R_{UR}^2)/(n-k)} \tag{4-39}$$

since the *SST* term is cancelled.

This is called the *R-squared* form of the F statistic.

Whereas the R-squared form of the F statistic is very useful for testing exclusion restrictions, it cannot be applied for testing all kinds of linear restrictions. For example, the F ratio (4-39) cannot be used when the model does not have intercept or when the functional form of the endogenous variable in the unrestricted model is not the same as in the restricted model.

Decision rule

The $F_{q,n-k}$ distribution is tabulated and available in statistical tables, where we look for the critical value ($F_{q,n-k}^{\alpha}$), which depends on α (the significance level), q (the *df* of the numerator), and *n-k*, (the *df* of the denominator). Taking into account the above, the *decision rule* is quite simple.

Decision rule

$$
\begin{array}{llll}
\text{If} & F \geq F^{\alpha}_{q,n-k} & \text{reject} & H_0 \\
\text{If} & F < F^{\alpha}_{q,n-k} & \text{not reject} & H_0
\end{array}
\qquad (4\text{-}40)
$$

Therefore, we reject H_0 in favor of H_1 at α when $F \geq F^{\alpha}_{q,n-k}$, as can be seen in figure 4.15. It is important to remark that as α decreases, $F^{\alpha}_{q,n-k}$ increases. If H_0 is rejected, then we say that $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ are *jointly statistically significant*, or just *jointly significant*, at the selected significance level.

This test alone does not allow us to say which of the variables has a partial effect on *y*; they may all affect *y* or only one may affect *y*. If H_0 is not rejected, then we say that $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ are jointly not statistically significant, or simply jointly not significant, which often justifies dropping them from the model. The *F* statistic is often useful for testing the exclusion of a group of variables when the variables in the group are highly correlated.

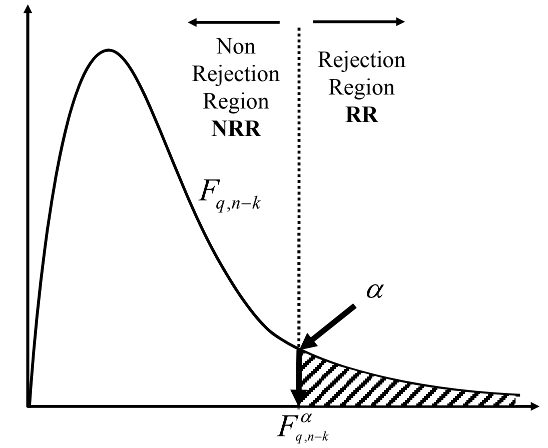


FIGURE 4.15. Rejection region and non rejection region using *F* distribution.

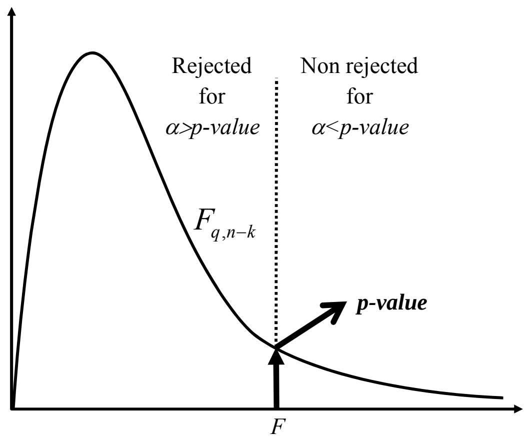

Rejected for α>*p-value*
Non rejected for α<*p-value*
$F_{q,n-k}$
p-value
F

FIGURE 4.16. *p-value* using *F* distribution.

In the *F* testing context, the *p*-value is defined as

$$p\text{-}value = \Pr(F > F' \mid H_0)$$

where *F* is the actual value of the test statistic and F' denotes a Snedecor's *F* random variable with (*q*,*n*-*k*) *df*.

The *p*-value still has the same interpretation as for *t* statistics. A small *p*-value is evidence against H_0, while a large *p*-value is not evidence against H_0. Once the *p*-value has been computed, the *F* test can be carried out at any significance level. In figure 4.16 this alternative approach is represented. As can be seen by observing the figure, the determination of the *p*-value is the inverse operation to find the value in the statistical tables for a given significance level. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, whereas the null hypothesis is not rejected when α<*p*-value.

EXAMPLE 4.10 Wage, experience, tenure and age

The following model has been built to analyze the determinant factors of wage:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 tenure + \beta_5 age + u$$

where *wage* is monthly earnings, *educ* is years of education, *exper is* years of work experience, *tenure* is years with current employer, and *age* is age in years.

The researcher is planning to exclude *tenure* from the model, since in many cases it is equal to experience, and also *age*, because it is highly correlated with experience. Is the exclusion of both variables acceptable?

The null and alternative hypotheses are the following:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : H_0 \text{ is not true}$$

The restricted model corresponding to this H_0 is

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + u$$

Using a sample consisting of 53 observations from workfile *wage2*, we have the following estimations for the unrestricted and for the restricted models:

$$\widehat{\ln(wage_i)} = 6.476 + 0.0658 educ_i + 0.0267 exper_i - 0.0094 tenure_i - 0.0209 age_i \quad RSS = 5.954$$

$$\widehat{\ln(wage_i)} = 6.157 + 0.0457 educ_i + 0.0121 exper_i \quad RSS = 6.250$$

The *F* ratio obtained is the following:

$$F = \frac{\left(RSS_R - RSS_{UR}\right)/q}{RSS_{UR}/(n-k)} = \frac{(6.250 - 5.954)/2}{5.954/48} = 1.193$$

Given that the *F* statistic is low, let us see what happens with a significance level of 0.10. In this case the degrees of freedom for the denominator are 48 (53 observations minus 5 estimated parameters). If we look in the *F* statistical table for 2 *df* in the numerator and 45 *df* in the denominator, we find $F_{2,48}^{0.10} \simeq F_{2,45}^{0.10}$ =2.42. As *F*<2.42, we do not reject H_0. If we do not reject H_0 for 0.10, we will not reject H_0 for 0.05 or 0.01, as can been in figure 4.17. Therefore, we cannot reject H_0 in favor of H_1. In other words *tenure* and *age* are not jointly significant.

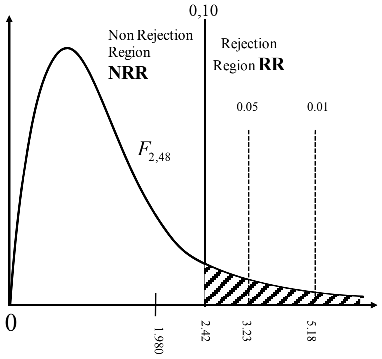


FIGURE 4.17. Example 4.10: Rejection region using *F* distribution (α values are from a $F_{2.40}$).

4.3.2 Model significance

Testing model significance, or overall significance, is a particular case of testing exclusion restrictions. Model significance means global significance of the model. One could think that the H_0 in this test is the following:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 \qquad (4\text{-}41)$$

However, this is not the adequate H_0 to test for the global significance of the model. If $\beta_2 = \beta_3 = \cdots = \beta_k = 0$, then the restricted model would be the following:

$$y = \beta_1 + u \qquad (4\text{-}42)$$

If we take expectations in (4-42), then we have

$$E(y)=\beta_1 \tag{4-43}$$

Thus, H_0 in (4-41) states not only that the explanatory variables have no influence on the endogenous variable, but also that the mean of the endogenous variable–for example, the consumption mean- is equal to 0.

Therefore, if we want to know whether the model is globally significant, the H_0 must be the following:

$$H_0:\beta_2=\beta_3=\cdots=\beta_k=0 \tag{4-44}$$

The corresponding restricted model given in (4-42) does not explain anything and, therefore, R_R^2 is equal to 0. Testing the H_0 given in (4-44) is very easy by using the *R-squared* form of the *F* statistic:

$$F=\frac{R^2/k}{(1-R^2)/(n-k)} \tag{4-45}$$

where R^2 is the R_{UR}^2, since only the unrestricted model needs to be estimated, because the R^2 of the model (4-42) – restricted model- is 0.

EXAMPLE 4.11 Salaries of CEOs

Consider the following equation to explain salaries of Chief Executive Officers (CEOs) as a function of annual firm sales, return on equity (*roe*, in percent form), and return on the firm's stock (*ros*, in percent form):

$$\ln(salary)=\beta_1+\beta_2\ln(sales)+\beta_3 roe+\beta_4 ros+u.$$

The question posed is whether the performance of the company (*sales*, *roe* and *ros*) is crucial to set the salaries of CEOs. To answer this question, we will carry out an overall significance test. The null and alternative hypotheses are the following:

$$H_0:\beta_2=\beta_3=\beta_4=0$$

$$H_1\text{: } H_0 \text{ is not true}$$

Table 4.9 shows an E-views complete output for *least square* (*ls*) using the filework *ceosal1*. At the bottom the "F-statistic" can be seen for overall test significance, as well as "Prob", which is the *p*-value corresponding to this statistic. In this case the *p*-value is equal to 0, that is to say, H_0 is rejected for all significance levels (See figure 4.18). Therefore, we can reject that the performance of a company has no influence on the salary of a CEO.

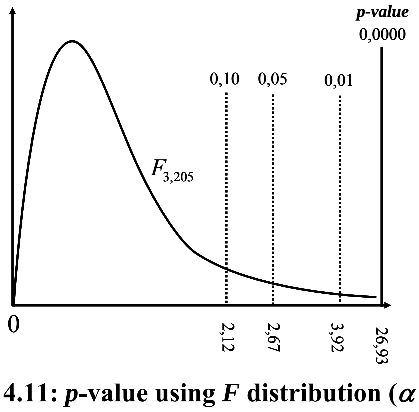


FIGURE 4.18. Example 4.11: *p*-value using *F* distribution (α values are for a $F_{3,140}$).

TABLE 4.9. Complete output from E-views in the example 4.11.

Dependent Variable:
LOG(SALARY)
Method: Least Squares
Date: 04/12/12 Time: 19:39
Sample: 1 209
Included observations: 209

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.311712	0.315433	13.66919	0.0000
LOG(SALES)	0.280315	0.03532	7.936426	0.0000
ROE	0.017417	0.004092	4.255977	0.0000
ROS	0.000242	0.000542	0.446022	0.6561

R-squared	0.282685	Mean dependent var	6.950386
Adjusted R-squared	0.272188	S.D. dependent var	0.566374
S.E. of regression	0.483185	Akaike info criterion	1.402118
Sum squared resid	47.86082	Schwarz criterion	1.466086
Log likelihood	-142.5213	**F-statistic**	**26.9293**
Durbin-Watson stat	2.033496	**Prob(F-statistic)**	**0.0000**

4.3.3 Testing other linear restrictions

So far, we have tested hypotheses with exclusion restrictions using the F statistic. But we can also test hypotheses with linear restrictions of any kind. Thus, in the same test we can combine exclusion restrictions, restrictions that impose determined values to the parameters and restrictions on linear combination of parameters.

Therefore, let us consider the following model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

and the null hypothesis:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_4 = 3 \\ \beta_5 = 0 \end{cases}$$

The restricted model corresponding to this null hypothesis is

$$(y - x_2 - 3x_4) = \beta_1 + \beta_3 (x_3 - x_2) + u$$

In the example 4.12, the null hypothesis consists of two restrictions: a linear combination of parameters and an exclusion restriction.

EXAMPLE 4.12 An additional restriction in the production function. (Continuation of example 4.7)

In the production function of Cobb-Douglas, we are going to test the following H_0 which has two restrictions:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_1 = 0 \end{cases}$$

$$H_1 : H_0 \text{ is not true}$$

In the first restriction we impose that there are constant returns to scale. In the second restriction that β_1, parameter linked to the total factor productivity is equal to 0.

Substituting the restriction of H_0 in the original model (*unrestricted model*), we have

$$\ln(output) = (1-\beta_3)\ln(labor) + \beta_3 \ln(capital) + u$$

Operating, we obtain the *restricted model*:

$$\ln(output / labor) = \beta_3 \ln(capital / labor) + u$$

In estimating the unrestricted and restricted models, we get RSS_R=3.1101 and RSS_{UR}=0.8516. Therefore, the *F ratio is*

$$F = \frac{\left(RSS_R - RSS_{UR}\right)/q}{RSS_{UR}/(n-k)} = \frac{(3.1101-0.8516)/2}{0.8516/(27-3)} = 13.551$$

There are two reasons for not using R^2 in this case. First, the restricted model has no intercept. Second, the regressand of the restricted model is different from the regressand of the unrestricted model.

Since the *F* value is relatively high, let us start by testing with a level of 1%. For α=0.01, $F_{2,24}^{0.01} = 5.61$. Given that *F*>5.61, we reject H_0 in favour of H_1. Therefore, we reject the joint hypotheses that there are constant returns to scale and that the parameter β_1 is equal to 0. If H_0 is rejected for α=0.01, it will also be rejected for levels of 5% and 10%.

4.3.4 Relation between *F* and *t* statistics

So far, we have seen how to use the *F* statistic to test several restrictions in the model, but it can be used to test a single restriction. In this case, we can choose between using the *F* statistic or the *t* statistic to carry out a two-tail test. The conclusions would, nevertheless, be exactly the same.

But, what is the relationship between an *F* with one degree of freedom in the numerator (to test a single restriction) and a *t*? It can be shown that

$$t_{n-k}^2 = F_{1,n-k} \tag{4-46}$$

This fact is illustrated in figure 4.19. We observe that the tail of the *F* splits into the two tails of the *t*. Hence, the two approaches lead to exactly the same outcome, provided that the alternative hypothesis is two-sided. However, the *t* statistic is more flexible for testing a single hypothesis, because it can be used to test H_0 against one-tail alternatives.

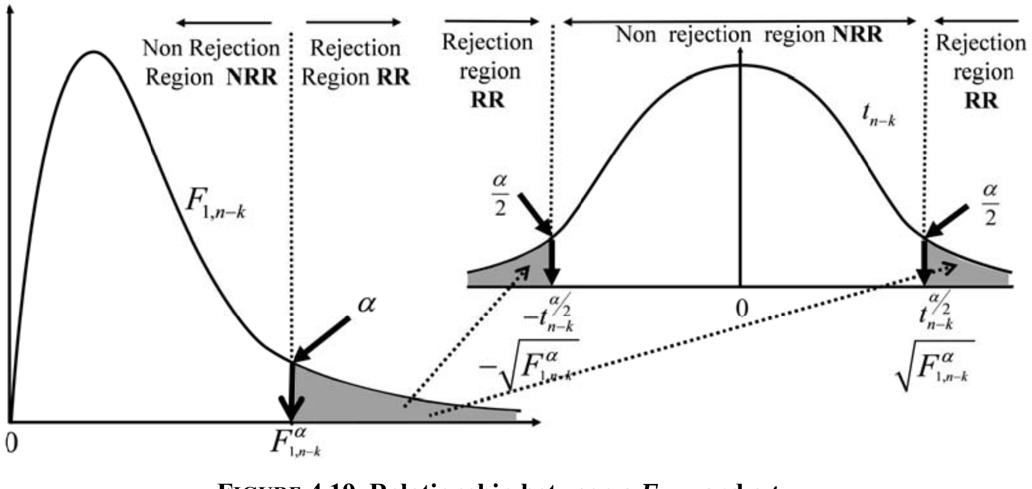


FIGURE 4.19. Relationship between a $F_{1,n-k}$ and a t_{n-k}.

Moreover, since the *t* statistics are also easier to obtain than the *F* statistics, there is no good reason for using an *F* statistic to test a hypothesis with a unique restriction.

4.4 Testing without normality

The normality of the *OLS* estimators depends crucially on the normality assumption of the disturbances. What happens if the disturbances do not have a normal distribution? We have seen that the disturbances under the Gauss-Markov assumptions, and consequently the *OLS* estimators are asymptotically normally distributed, i.e. approximately normally distributed.

If the disturbances are not normal, the *t* statistic will only have an *approximate t* distribution rather than an *exact* one. As it can be seen in the *t* student table, for a sample size of 60 observations the critical points are practically equal to the standard normal distribution.

Similarly, if the disturbances are not normal, the *F* statistic will only have an *approximate F* distribution rather than an *exact* one, when the sample size is large enough and the Gauss-Markov assumptions are fulfilled. Therefore, we can use the *F* statistic to test linear restrictions in linear models as an approximate test.

There are other asymptotic tests (the likelihood ratio, Lagrange multiplier and Wald tests) based on the likelihood functions that can be used in testing linear restriction if the disturbances are non-normally distributed. These three can also be applied when a) the restrictions are nonlinear; and b) the model is nonlinear in the parameters. For non-linear restrictions, in linear and non-linear models, the most widely used test is the Wald test.

For testing the assumptions of the model (for example, homoskedasticity and no autocorrelation) the Lagrange multiplier (*LM*) test is usually applied. In the application of the *LM* test, an *auxiliary regression* is often run. The name of auxiliary regression means that the coefficients are not of direct interest: only the R^2 is retained. In an auxiliary regression the regressand is usually the residuals (or functions of the residuals), obtained in the OLS estimation of the original model, while the regressors are often the regressors (and/or functions of them) of the original model.

4.5 Prediction

In this section two types of prediction will be examined: point and interval prediction.

4.5.1 Point prediction

Obtaining a point prediction does not pose any special problems, since it is a simple extrapolation operation in the context of descriptive methods.

Let $x_2^0, x_3^0, \cdots, x_k^0$ denote the particular values in each of the *k* regressors for prediction; these may or may not correspond to an actual data point in our sample. If we substitute these values in the multiple regression model, we have

$$y^0 = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 + u^0 = \theta^0 + u^0 \qquad (4\text{-}47)$$

Therefore, the expected, or mean, value of y is given by

$$E(y^0) = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \qquad (4\text{-}48)$$

The point prediction is obtained straightaway by replacing the parameters of (4-48) by the corresponding OLS estimators:

$$\hat{\theta}^0 = \hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0 \tag{4-49}$$

To obtain (4-49) we did not need any assumption. But, if we adopt the assumptions 1 to 6, we will immediately find that that $\hat{\theta}^0$ is an unbiased predictor of θ^0 :

$$E\left[\hat{\theta}^0\right] = E\left[\hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0\right] = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + \ldots + \beta_k x_k^0 = \theta^0 \tag{4-50}$$

On the other hand, adopting the Gauss Markov assumptions (1 to 8), it can be proved that this point predictor is the best linear unbiased estimator (BLUE).

We have a point prediction for θ^0, but, what is the point prediction for y^0? To answer this question, we have to predict u_0. As the error is not observable, the best predictor for u^0 is its expected value, which is 0. Therefore,

$$\hat{y}^0 = \hat{\theta}^0 \tag{4-51}$$

4.5.2 Interval prediction

Point predictions made with an econometric model will in general not coincide with the observed values due to the uncertainty surrounding economic phenomena.

The first source of uncertainty is that we cannot use the population regression function because we do not know the parameters β's. Instead we have to use the sample regression function. The *confidence interval for the expected value* – i.e. for θ^0 - which will examine next, includes only this type of uncertainty.

The second source of uncertainty is that in an econometric model, in addition to the systematic part, there is a disturbance which is not observable. The *prediction interval for an individual value* – i.e. for y^0-, which will be discussed later on includes both the uncertainty arising from the estimation as well as the disturbance term.

A third source of uncertainty may come from the fact of not knowing exactly what values the explanatory variables will take for the prediction we want to make. This third source of uncertainty, which is not addressed here, complicates calculations for the construction of intervals.

Confidence interval for the expected value

If we are predicting the expected value of y, which is θ^0, then the prediction error $\hat{e}_1^0$ will be $\hat{e}_1^0 = \theta^0 - \hat{\theta}^0$. According to (4-50), the expected prediction error is zero. Under the assumptions of the *CLM*,

$$\frac{\hat{e}_1^0}{se(\hat{\theta}^0)} = \frac{\theta^0 - \hat{\theta}^0}{se(\hat{\theta}^0)} \sim t_{n-k}$$

Therefore, we can write that

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\theta^0 - \hat{\theta}^0}{se(\hat{\theta}^0)} \le t_{n-k}^{\alpha/2}\right] = 1 - \alpha$$

Operating, we can construct a (1-α)% *confidence interval* (*CI*) for θ^0 with the following structure:

$$\Pr\left[\hat{\theta}^0 - se(\hat{\theta}^0)\times t_{n-k}^{\alpha/2} \le \theta^0 \le \hat{\theta}^0 + se(\hat{\theta}^0)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha \qquad (4\text{-}52)$$

To obtain a *CI* for θ^0, we need to know the standard error ($se(\hat{\theta}_0)$) for $\hat{\theta}^0$. In any case, there is an easy way to calculate it. Thus, solving (4-48) for β_1 we find that $\beta_1 = \theta^0 - \beta_2 x_2^0 - \beta_3 x_3^0 - ... - \beta_k x_k^0$. Plugging this into the equation (4-47), we obtain

$$y = \theta^0 + \beta_2(x_2 - x_2^0) + \beta_3(x_3 - x_3^0) + \ldots + \beta_k(x_k - x_k^0) + u \qquad (4\text{-}53)$$

Applying OLS to (4-53), in addition to the point prediction, we obtain $se(\hat{\theta}^0)$ which is the standard error corresponding to the *intercept* in this regression. The previous method allows us to put a *CI* around the OLS estimate of *E*(*y*), for any values of the *x*′s.

Prediction interval for an individual value

We are now going to construct an interval for y^0, usually called *prediction interval for an individual value,* or for short, *prediction interval*. According to (4-47), y^0 has two components:

$$y^0 = \theta^0 + u^0 \qquad (4\text{-}54)$$

The *interval for the expected value* built before is a confidence interval around θ^0 wcich is a combination of the parameters. In contrast, the interval for y^0 is random, because one of its components, u^0, is random. Therefore, the interval for y^0 is a probabilistic interval and not a confidence interval. The mechanics for obtaining it are the same, but bear in mind that now we are going to consider that the set $x_2^0, x_3^0, \cdots, x_k^0$ vis outside from of the sample used to estimate the regression.

The *prediction error* ($\hat{e}_2^0$) in using $\hat{y}^0$ to predict y^0 is

$$\hat{e}_2^0 = y^0 - \hat{y}^0 = \theta^0 + u^0 - \hat{y}^0 \qquad (4\text{-}55)$$

Taking into account (4-51) and (4-50), and that $E(u^0)=0$, then the expected prediction error is zero. In finding the variance of $\hat{e}_2^0$, it must be taken into account that u^0 is uncorrelated with $\hat{y}^0$ because $x_2^0, x_3^0, \cdots, x_k^0$ is not in the sample.

Therefore, the *variance of the prediction error* (conditional on the *x*′s) is the sum of the variances:

$$Var(\hat{e}_2^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2 \qquad (4\text{-}56)$$

There are two sources of variation in $\hat{e}_2^0$:

1. The sampling error in $\hat{y}^0$, which arises because we have estimated the β_j*'s*.
2. The ignorance of the unobserved factors that affect *y*, which is reflected in σ^2.

Under the *CLM* assumptions, $\hat{e}_2^0$ is also normally distributed. Using the unbiased estimator of σ^2 and taking into account that $var(\hat{y}^0) = var(\hat{\theta}^0)$, we can define the standard error (*se*) of $\hat{e}_2^0$ as

$$se(\hat{e}_2^0)=\left\{\left[se(\hat{\theta}^0)\right]^2+\hat{\sigma}^2\right\}^{\frac{1}{2}} \tag{4-57}$$

Usually $\hat{\sigma}^2$ is larger than $\left[se(\hat{\theta}^0)\right]^2$. Under the assumptions of the *CLM*,

$$\frac{\hat{e}_2^0}{se(\hat{e}_2^0)}\sim t_{n-k} \tag{4-58}$$

Therefore, we can write that

$$\Pr\left[-t_{n-k}^{\alpha/2}\leq\frac{\hat{e}_2^0}{se(\hat{e}_2^0)}\leq t_{n-k}^{\alpha/2}\right]=1-\alpha \tag{4-59}$$

Plugging in $\hat{e}_2^0=y^0-\hat{y}^0$ into (4-59) and rearranging it gives a (1-α)% *prediction interval* for y^0:

$$\Pr\left[\hat{y}^0-se(\hat{e}_2^0)\times t_{n-k}^{\alpha/2}\leq y^0\leq\hat{y}^0+se(\hat{e}_2^0)\times t_{n-k}^{\alpha/2}\right]=1-\alpha \tag{4-60}$$

Example 4. 13 What is the expected score in the final exam with 7 marks in the first short exam?

The following model has been estimated to compare the marks in the final exam (*finalmrk*) and in the first short exam (*shortex1*) of Econometrics:

$$\widehat{finalmrk_i}=\underset{(0.715)}{4.155}+\underset{(0.123)}{0.491}shortex1_i$$

$$\hat{\sigma}=1.649\quad R^2=0.533\quad n=16$$

To estimate the expected final mark for a student with *shortex1*^{0}=7 mark in the first short exam, the following model, according to (4-53), was estimated:

$$\widehat{finalmrk_i}=\underset{(0.497)}{7.593}+\underset{(0.123)}{0.491}\left(shortex1_i-7\right)$$

$$\hat{\sigma}=1.649\quad R^2=0.533\quad n=16$$

The point prediction for *shortex1*^{0}=7 is $\hat{\theta}_0$=7.593 and the lower and upper bounds of a 95% *CI* respectively are given by

$$\underline{\theta}^0=\hat{\theta}^0-se(\hat{\theta}^0)\times t_{14}^{0.05/2}=7.593-0.497\times 2.14=6.5$$

$$\bar{\theta}^0=\hat{\theta}^0+se(\hat{\theta}^0)\times t_{14}^{0.05/2}=7.593+0.497\times 2.14=8.7$$

Therefore, the student will have a 95% confidence of obtaining on average a final mark located between 6.5 and 8.7.

The point prediction could be also obtained from the first estimated equation:

$$\widehat{finalmrk}=4.155+0.491\times 7=7.593$$

Now, we are going to estimate a 95% probability interval for the individual value. The *se* of $\hat{e}_2^0$ is equal

$$se(\hat{e}_2^0)=\left\{\left[se(\hat{y}^0)\right]^2+\hat{\sigma}^2\right\}^{\frac{1}{2}}=\sqrt{0.497^2+1.649^2}=1.722$$

where 1.649 is the "S. E. of regression" obtained from the E-views output directly.

The lower and upper bounds of a 95% *probability interval* respectively are given by

$$\underline{y}^0=\hat{y}^0-se(\hat{e}_2^0)\times t_{14}^{0.025}=7.593-1.722\times 2.14=3.7$$

$$\bar{y}^0=\hat{y}^0+se(\hat{e}_2^0)\times t_{14}^{0.025}=7.593+1.722\times 2.14=11.3$$

You must take into account that this *probability interval* is quite large because the size of the sample is very small.

EXAMPLE 4.14 Predicting the salary of CEOs

Using data on the most important US companies taken from Forbes (workfile *ceoforbes*), the following equation has been estimated to explain salaries (including bonuses) earned yearly (thousands of dollars) in 1999 by the CEOs of these companies:

$$\widehat{salary}_i = \underset{(104)}{1381} + \underset{(0.0013)}{0.008377}\, assets_i + \underset{(8.671)}{32.508}\, tenure_i + \underset{(0.0538)}{0.2352}\, profits_i$$

$$\hat{\sigma} = 1506 \quad R^2 = 0.2404 \quad n = 447$$

where *assets* are total assets of firm in millions of dollars, *tenure* is number of years as CEO in the company, and *profits* are in millions of dollars.

In Table 4.10 descriptive measures of explanatory variables of the model on CEOs salaries appear.

TABLE 4.10. Descriptive measures of variables of the model on CEOs salary.

	assets	*tenure*	*profits*
Mean	27054	7.8	700
Median	7811	5.0	333
Maximum	668641	60.0	22071
Minimum	718	0.0	-2669
Observations	447	447	447

The predicted salaries and the corresponding $se(\hat{\theta}_0)$ for selected values (maximum, mean, median and minimum), using a model as (4-53), appear in table 4.11.

TABLE 4.11. Predictions for selected values.

	Prediction $\hat{\theta}_0$	*Std. Error* $se(\hat{\theta}_0)$
Mean values	2026	71
Median value	1688	78
Maximum values	14124	1110
Minimum values	760	195

4.5.3 Predicting *y* in a ln(*y*) model

Consider the model in logs:

$$\ln(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (4\text{-}61)$$

Obtaining OLS estimates, we predict ln(*y*) as

$$\widehat{\ln(y)} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \qquad (4\text{-}62)$$

Applying exponentiation to (4-62), we obtain the prediction value

$$\tilde{y} = \exp(\widehat{\ln(y)}) = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \qquad (4\text{-}63)$$

However, this prediction is biased and inconsistent because it will systematically *underestimate* the expected value of *y*. Let us see why. If we apply exponentiation in (4-61), we have

$$y = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(u) \qquad (4\text{-}64)$$

Before taking expectation in (4-64), we must take into account that if $u \sim N(0,\sigma^2)$, then $E(\exp(u)) = \exp\left(\frac{\sigma^2}{2}\right)$. Therefore, under the *CLM* assumptions 1 through 9, we have

$$E(y) = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(\sigma^2 / 2) \qquad (4\text{-}65)$$

Taking as a reference (4-65), the adequate predictor of y is

$$\hat{y} = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \times \exp(\hat{\sigma}^2 / 2) = \tilde{y} \times \exp(\hat{\sigma}^2 / 2) \tag{4-66}$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2.

It is important to remark that although $\hat{y}$ is a biased predictor, it is consistent, while $\tilde{y}$ is biased and inconsistent

EXAMPLE 4.15 Predicting the salary of CEOs with a log model (continuation 4.14)

Using the same data as in example 4.14, the following model was estimated:

$$\widehat{\ln(salary_i)} = \underset{(0.210)}{5.5168} + \underset{(0.0232)}{0.1885}\ln(assets_i) + \underset{(0.0032)}{0.0125}\,tenure_i + \underset{(0.0000195)}{0.00007}\,profits_i$$

$\hat{\sigma}$ =0.5499 R^2=0.2608 n=447

Salary and *assets* are taken in natural logs, while *profits* are in levels because some observations are negative and thus not possible to take logs.

First, we are going to calculate the inconsistent prediction, according to (4-63) for a CEO working in a corporation with *assets*=10000, *tenure*=10 years and *profits*=1000:

$$\widetilde{salary}_i = \exp(\widehat{\ln(salary_i)})$$

$$= \exp(5.5168 + 0.1885\ln(10000) + 0.0125 \times 10 + 0.00007 \times 1000) = 1716$$

Using (4-66), we obtain a consistent prediction:

$$\widehat{salary} = \exp(0.5499^2 / 2) \times 1716 = 1996$$

4.5.4 Forecast evaluation and dynamic prediction

In this section we will compare predictions made using an econometric model with the actual values in order to evaluate the predictive ability of the model. We will also examine the dynamic prediction in models in which lagged endogenous variables are included as regressors.

Forecast evaluation statistics

Suppose that the sample forecast is i=n+1, n+2,…, n+h, and denote the actual and forecasted value in period i as y_i and $\hat{y}_i$, respectively. Now, we present some of the more common statistics used for forecast evaluation.

Mean absolute error (MAE)

The *MAE* is defined as the average of the absolute values of the errors:

$$MAE = \frac{\sum_{i=n+1}^{n+h} \left| \hat{y}_i - y_i \right|}{h} \tag{4-67}$$

Absolute values are taken so that positive errors are compensated by the negative ones.

Mean absolute percentage error (MAPE),

$$MAPE = \frac{\sum_{i=n+1}^{n+h} \frac{\left| \hat{y}_i - y_i \right|}{y_i}}{h} \times 100 \tag{4-68}$$

Root of the mean squared error (RMSE)

This statistic is defined as the square root of the mean of the squared error:

$$RMSE=\sqrt{\frac{\sum_{i=n+1}^{n+h}\left(\hat{y}_i-y_i\right)^2}{h}} \tag{4-69}$$

As the errors are squared, the compensation between positive and negative errors are avoided. It is important to remark that the *MSE* places a greater penalty on large forecast errors than the *MAE*.

Theil Inequality Coefficient (U)

This coefficient is defined as follows:

$$U=\frac{\sqrt{\frac{\sum_{i=n+1}^{n+h}\left(\hat{y}_i-y_i\right)^2}{h}}}{\sqrt{\frac{\sum_{i=n+1}^{n+h}\hat{y}_i^2}{h}}+\sqrt{\frac{\sum_{i=n+1}^{n+h}y_i^2}{h}}} \tag{4-70}$$

The smaller *U* is, the more accurate are the predictions. The scaling of *U* is such that it will always lie between 0 and 1. If *U*=0, then $y_i=\hat{y}_i$, for all forecasts; if *U*=1 the predictive performance is as bad as it can be. Theil's *U* statistic can be rescaled and decomposed into three proportions: bias, variance and covariance. Of course the sum of these three proportions is 1. The interpretation of these three proportions is as follows:

1) The *bias* reflects systematic errors. Whatever the value of *U*, we would hope that the bias is close to 0. A large bias suggests a systematic over or under prediction.
2) The *variance* also reflects systematic errors. The size of this proportion is an indication of the inability of the forecasts to replicate the variability of the variable to be forecasted.
3) The *covariance* measures unsystematic errors. Ideally, this should have the highest proportion of Theil inequality.

In addition of the coefficient defined in (4-70), Theil proposed other coefficients for forecast evaluation.

Dynamic prediction

Let the following model be given:

$$y_t=\beta_1+\beta_2x_t+\beta_3y_{t-1}+u_t \tag{4-71}$$

Suppose that the sample forecast is *i*=*n*+1,…,*i*=*n*+h, and denote the actual and forecasted value in period *i* as y_i and $\hat{y}_i$, respectively. The forecast for the period *n*+1 is

$$\hat{y}_{n+1}=\hat{\beta}_1+\hat{\beta}_2x_{n+1}+\hat{\beta}_3y_n \tag{4-72}$$

As we can see for the prediction, we use the observed value of *y* (y_n) because it is inside the sample used in the estimation. For the remainder of the forecast periods we use the recursively computed forecast of the lagged value of the dependent variable (dynamic prediction), that is to say,

$$\hat{y}_{n+i} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+i} + \hat{\beta}_3 \hat{y}_{n-1+i} \qquad i = 2,3,\cdots,h \qquad (4\text{-}73)$$

Thus, from period n+2 to n+h the forecast carried out in a period is used to forecast the endogenous variable in the following period.

Exercises

Exercise 4.1 To explain the housing price in an American town, the following model is formulated:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

where *rooms* is the number of rooms in the house, *lowstat* is the percentage of people of "lower status" in the area and *crime* is crimes committed per capita in the area. Prices of houses are measured in dollars.

Using the data in *hprice2*, the following model has been estimated:

$$\widehat{price} = \underset{(8022)}{-15694} + \underset{(1211)}{6788} rooms \underset{(81)}{-268} lowstat - \underset{(960)}{3854} crime$$

$$R^2 = 0.771 \quad n = 55$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the meaning of the coefficients $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$.

b) Does the percentage of people of "lower status" have a negative influence on the price of houses in that area?

c) Does the number of rooms have a positive influence on the price of houses?

Exercise 4.2 Consider the following model:

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + \beta_4 punder5 + u$$

where *fruit* is expenditure in fruit, *inc* is disposable income of a household, *hhsize* is the number of household members and *punder5* is the proportion of children under five in the household.

Using the data in workfile *demand*, the following model has been estimated:

$$\widehat{\ln(fruit)} = \underset{(3.701)}{-9.768} + \underset{(0.512)}{2.005} \ln(inc) - \underset{(0.179)}{1.205} hhsize \underset{(0.013)}{-0.0179} punder5$$

$$R^2 = 0.728 \quad n = 40$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the meaning of the coefficients $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$.

b) Does the number of household members have a statistically significant effect on the expenditure in fruit?

c) Is the proportion of children under five in the household a factor that has a negative influence on the expenditure of fruit?

d) Is fruit a luxury good?

Exercise 4.3 (Continuation of exercise 2.5). Given the model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad \text{i} = 1,2,\ldots,n$$

the following results have been obtained with a sample size of 11 observations:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

(Remember that $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$)

a) Build a statistic to test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$.

b) Test the hypothesis of question *a)* when $EB = 2F^2$.

c) Test the hypothesis of question *a)* when $EB = F^2$.

Exercise 4.4 The following model has been formulated to explain the spending on food (*food*):

$$food = \beta_1 + \beta_2 inc + \beta_3 rpfood + u$$

where *inc* is disposable income and *rpfood* is the relative price index of food compared to other consumer products.

Taking a sample of observations for 20 successive years, the following results are obtained:

$$\widehat{food_i} = \underset{(4.92)}{1.40} + \underset{(0.01)}{0.126}\, inc_i - \underset{(0.07)}{0.036}\, rpfood_i$$

$$R^2 = 0.996; \qquad \sum \hat{u}_t^2 = 0.196$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the null hypothesis that the coefficient of *rpfood* is less than 0.

b) Obtain a confidence interval of 95% for the marginal propensity to consume food in relation to income.

c) Test the joint significance of the model.

Exercise 4.5 The following demand function for rental housing is formulated:

$$\ln(srenhous_i) = \beta_1 + \beta_2 \ln(prenhous_i) + \beta_3 \ln(inc_i) + \varepsilon_i$$

where *srenhous* is spending on rental housing, *prenhous* is the rental price, and *inc* is disposable income.

Using a sample of 403 observations, we obtain the following results:

$$\ln(srenhous_i) = 10 - 0.7\ln\left(prenhous_i\right) + 0.9\ln\left(inc_i\right)$$

$$R^2 = 0.39 \qquad \text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 0.09 & 0.085 \\ 0 & 0.085 & 0.09 \end{bmatrix}$$

a) Interpret the coefficients on ln(*prenhous*) and ln(*inc*).

b) Using a 0.01 significance level, test the null hypothesis that $\beta_2 = \beta_3 = 0$.

c) Test the null hypothesis that $\beta_2 = 0$, against the alternative that $\beta_2 < 0$.

d) Test the null hypothesis that $\beta_3 = 1$ against the alternative that $\beta_3 \neq 1$.

e) Test the null hypothesis that a simultaneous increase in housing prices and income has no proportional effect on housing demand.

Exercise 4.6 The following estimated models corresponding to average cost (*ac*) functions have been obtained, using a sample of 30 firms:

$$\widehat{ac_i} = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, qty_i$$

$$R^2 = 0.838 \qquad RSS = 8090 \qquad (1)$$

$$\widehat{ac_i} = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, qty_i + \underset{(11.61)}{26.73}\, qty_i^2 - \underset{(1.22)}{1.40}\, qty_i^3$$

$$R^2 = 0.978 \quad RSS = 1097 \qquad (2)$$

where *ac* is the average cost and *qty* is the quantity produced.

(The numbers in parentheses are standard errors of estimators.)

a) Test whether the quadratic and cubic terms of the quantity produced are significant in determining the average cost.

b) Test the overall significance in the model 2.

Exercise 4.7 Using a sample of 35 observations, the following models have been estimated to explain expenditures on coffee:

$$\widehat{\ln(coffee)} = 21.32 + \underset{(0.01)}{0.11}\ln(inc) - \underset{(0.23)}{1.33}\ln(cprice) + 1.35\ln(tprice)$$

$$R^2 = 0.905 \qquad RSS = 254 \qquad (1)$$

$$\widehat{\ln(coffee)} = 19.9 + \underset{(0.02)}{0.14}\ln(inc) - \underset{(0.21)}{1.42}\ln(cprice)$$

$$RSS = 529 \qquad (2)$$

where *inc* is disposable income, *cprice* is coffee price and *tprice* is tea price.

(The numbers in parentheses are standard errors of estimators.)

a) Test the overall significance of model (1)

b) The standard error of ln(*tprice*) is missing in model (1), can you calculate it?

c) Test whether the price of tea is statistically significant.

d) How would you test the assumption that the price elasticity of coffee is equal but opposite to the price elasticity of tea? Detail the procedure.

Exercise 4.8 The following model has been formulated to analyse the determinants of air quality (*airqual*) in 30 Standard Metropolitan Statistical Areas (SMSA) of California:

$$airqual = \beta_1 + \beta_2 popln + \beta_3 medincm + \beta_4 poverty + \beta_5 fueoil + \beta_6 valadd + u$$

where *airqual* is weight in µg/m^3 of suspended particular matter, *popln* is population in thousands, *medincm* is medium per capita income in dollars, *poverty* is the percentage of families with income less than poverty levels, *fueloil* is thousands of barrels of fuel oil consumed in industrial manufacturing, and *valadd* is value added by industrial manufactures in 1972 in thousands of dollars.

Using the data in workfile *airqualy*, the above model has been estimated:

$$\widehat{airqual_i} = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$
$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011}\, valadd_i$$

$$R^2=0.415 \quad n=30$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the coefficients on *medincm*, *poverty* and *valadd*

b) Are the slope coefficients individually significant at 10%?

c) Test the joint significance of *fueloil* and *valadd*, knowing that

$$\widehat{airqual_i} = \underset{(10.41)}{97.67} + \underset{(0.020)}{0.0566}\,popln_i - \underset{(0.0039)}{0.0102}\,medincm_i - \underset{(0.0078)}{0.0174}\,poverty_i$$

$$R^2 = 0.339 \quad n = 30$$

d) If you omit the variable *poverty* in the first model, the following results are obtained:

$$\widehat{airqual_i} = \underset{(10.02)}{82.98_i} + \underset{(0.031)}{0.0523}\,popln_i - \underset{(0.0055)}{0.0097}\,medincm_i$$

$$-\underset{(0.0017)}{0.00063}\,fueoil_i - \underset{(0.0028)}{0.00037}\,valadd_i$$

$$R^2 = 0.218 \quad n = 30$$

Are the slope coefficients individually significant at 10% in the new model? Do you consider these results to be reasonable in comparison with those obtained in part *b*).

Comparing the R^2 of the two estimated models, what is the role played by *poverty* in determining air quality?

e) If you regress *airqual* using as regressors only the intercept and *poverty*, you will obtain that R^2=0.037. Do you consider this value to be reasonable taking into account the results obtained in part *d*)?

Exercise 4.9 With a sample of 39 observations, the following production functions by *OLS* was estimated:

$$\widehat{output_t} = \hat{\alpha} labor_t^{1.30} capital_t^{0.32} \exp(0.0055 trend_t) \qquad R^{2} = 0.9945$$

$$\widehat{output_t} = \hat{\beta} labor_t^{1.41} capital_t^{0.47} \qquad R^{2} = 0.9937$$

$$\widehat{output_i} = \hat{\gamma} \exp(0.0055 trend_t) \qquad R^{2} = 0.9549$$

a) Test the joint significance of *labor* and *capital.*

b) Test the significance of the coefficient of the variable *trend.*

c) Identify the statistical assumptions under which the test carried out in the two previous sections are correct. A further question: Specify the population model of the first of the three previous specifications.

Exercise 4.10 A researcher has developed the following model:

$$y = \beta_1 + \beta_2\, x_2 + \beta_3 x_3 + u$$

Using a sample of 43 observations, the following results were obtained:

$$\hat{y}_i = -\ 0.06 + 1.44\, x_{1i} - 0.48\, x_{2i}$$

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 0.1011 & -0.0007 & -0.0005 \\ & 0.0231 & -0.0162 \\ & & 0.0122 \end{bmatrix}$$

$$\sum y_i^2 = 444 \qquad \sum \hat{y}_i^2 = 424.92$$

a) Test that the intercept is less than 0.

b) Test that β_2=2.

c) Test the null hypothesis that β_2+3β_3=0.

Exercise 4.11 Given the function of production

$$q \; = ak^{\alpha} l^{\beta} \exp(u)$$

and using data from the Spanish economy over the past 20 years, the following results were obtained:

$$\widehat{\ln(q_i)} = 0.15 + 0.73 \ln(k_i) + 0.47 \ln(l_i)$$

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4129 & -95 & -266 \\ -95 & 3 & 5 \\ -266 & 5 & 19 \end{bmatrix} \qquad RSS = 0.017$$

a) Test the individual significance of the coefficients on *k* and *l.*

b) Test whether the parameter α is significantly different from 1.

c) Test whether there are increasing returns to scale.

Exercise 4.12 Let the following multiple regression model be:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$$

With a sample of 33 observations, this model is estimated by *OLS*, obtaining the following results:

$$\hat{y}_i = 12.7 + 14.2 x_{1i} + 2.1 x_{2i}$$

$$\hat{\sigma}^2 \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4.1 & -0.95 & -0.266 \\ -0.95 & 3.8 & 0.5 \\ -0.266 & 0.5 & 1.9 \end{bmatrix}$$

a) Test the null hypothesis α_0= α_1.

b) Test whether $\alpha_1 / \alpha_2 = 7$.

c) Are the coefficients α_0, α_1, y α_2 individually significant?

Exercise 4.13 Using a sample of 30 companies, the following cost functions have been estimated:

$$a)\ \widehat{cost_i} = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72} x_i \qquad R^2 = 0.838 \quad \bar{R}^2 = 0.829 \quad RSS = 8090$$

$$b)\ \widehat{cost_i} = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39} x_i + \underset{(11.61)}{26.73} x_i^2 - \underset{(1.22)}{1.40} x_i^3 \quad R^2 = 0.978 \quad \bar{R}^2 = 0.974 \quad RSS = 1097$$

where *cost* is the average cost and *x* is the quantity produced.

(The numbers in parentheses are standard errors of estimators.)

a) Which of the two models would you choose? What would be the criteria?

b) Test whether the quadratic and cubic terms of the quantity produced are significant in determining the average cost.

c) Test the overall significance of the model *b).*

Exercise 4.14 A researcher formulates the following model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Using a sample of 13 observations the following results are obtained:

$$\begin{aligned} \hat{y}_i &= 1.00 - 1.82x_{2i} + 0.36x_{3i} \\ R^2 &= 0.50 \quad n = 13 \\ \text{var}(\hat{\boldsymbol{\beta}}) &= \begin{bmatrix} 0.25 & -0.01 & 0.04 \\ -0.01 & 0.16 & -0.15 \\ 0.04 & -0.15 & 0.81 \end{bmatrix} \end{aligned} \qquad (1)$$

a) Test the null hypothesis that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 < 0$.

b) Test the null hypothesis that $\beta_2 + \beta_3 = -1$ against the alternative hypothesis that $\beta_2 + \beta_3 \neq -1$, with a significance level of 5%.

c) Is the whole model significant?

d) Assuming that the variables in the estimated model are measured in natural logarithms, what is the interpretation of the coefficient for x_3?

Exercise 4.15 With a sample of 50 automotive companies the following production functions were estimated taking the gross value added of the automobile production (*gva*) as the endogenous variable and labor input (*labor*) and capital input (*capital*) as explanatory variables.

$$1)\quad \begin{aligned} \widehat{\ln(gva_i)} &= 3.87 + \underset{(0.11)}{0.80}\ln(labor_i) + \underset{(0.24)}{1.24}\ln(capital_i) \\ RSS &= 254 \quad R^2 = 0.75 \quad \bar{R}^2 = 0.72 \end{aligned},$$

$$2)\quad \begin{aligned} &\widehat{\ln(gva_i)} = 19.9 + 1.04\ln(capital_i) \\ &RSS = 529 \quad R^2 = 0.84, \bar{R}^2 = 0.81 \end{aligned}$$

$$3)\quad \begin{aligned} &\widehat{\ln(gva\,/\,labor_i)} = 15.2 + 0.87\ln(capital_i\,/\,labor_i) \\ &RSS = 380 \end{aligned}$$

(The numbers in parentheses are standard errors of estimators.)

a) Test the joint significance of both factors in the production function.

b) Test whether labor has a significant positive influence on the gross value added of automobile production.

c) Test the hypothesis of constant returns to scale. Explain your answer.

Exercise 4.16 With a sample of 35 annual observations two demand functions of Rioja wine have been estimated. The endogenous variable is spending on Rioja reserve wine (*wine*) and the explanatory variables are disposable income (*inc*), the average price of a bottle of Rioja reserve wine (*pwinrioj*) and the average price of a bottle of Ribera Duero reserve wine (*pwinduer*). The results are as follows:

$$\widehat{\ln(vino_i)} = 21.32 + \underset{(0.01)}{0.11}\ln(renta_i) - \underset{(0.23)}{1.33}\ln(pvinrioj_i) + \underset{(0.233)}{1.35}\ln(pvinduer_i)$$

$$R^2 = 0.905 \quad RSS = 254$$

$$\widehat{\ln(vino_i)} = 19.9 + \underset{(0.02)}{0.14}\ln(renta_i) - \underset{(0.21)}{1.42}\ln(pvinrioj_i)$$

$$RSS = 529$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the joint significance of the first model.

b) Test whether the price of wine from Ribera del Duero has a significant influence, using two statistics that do not use the same information. Show that both procedures are equivalent.

c) How would you test the hypothesis that the price elasticity of Rioja wine is the same but with an opposite sign to the price elasticity of Ribera del Duero wine? Detail the procedure to follow.

Exercise 4.17 To analyze the demand for Ceylon tea (*teceil*) the following econometric model is formulated:

$$\ln(teceil) = \beta_1 + \beta_2 \ln(inc) + \beta_3 \ln(pteceil) + \beta_4 \ln(pteind) + \beta_5 \ln(pcobras) + u$$

where *inc* is the disposable income, *pteceil* the price of tea in Ceylon, *pteind* is the price of tea in India and *pcobras* is the price of Brazilian coffee.

With a sample of 22 observations the following estimates were made:

$$\widehat{\ln(teceil_i)} = 2.83 + \underset{(0.17)}{0.25} \ln(inc_i) - \underset{(0.98)}{1.48} \ln(pteceil_i)$$

$$+ \underset{(0.69)}{1.18} \ln(pteind_i) + \underset{(0.16)}{0.19} \ln(pcofbras_i)$$

RSS=0.4277

$$\widehat{\ln(teceil_i \times pteceil)} = 0.74 + \underset{(0.16)}{0.26} \ln(inc_i) + \underset{(0.15)}{0.20} \ln(pcofbras_i)$$

RSS=0.6788

(The numbers in parentheses are standard errors of the estimators.)

a) Test the significance of disposable income.

b) Test the hypothesis that $\beta_3 = -1$ y $\beta_4 = 0$, and explain the procedure applied.

c) If instead of having information on *RSS*, only R^2 was known for each model, how would you proceed to test the hypothesis of section *b)*?

Exercise 4.18 The following fitted models are obtained to explain the deaths of children under 5 years per 1000 live births (*deathu5*) using a sample of 64 countries.

$$1)\ \widehat{deathun5_i} = 263.64 - \underset{(0.0019)}{0.0056} inc_i + \underset{(0.21)}{2.23} fertrate_i; \qquad R^2 = 0.7077$$

$$2)\ \widehat{deathun5_i} = 168.31 - \underset{(0.0018)}{0.0055} inc_i + \underset{(0.25)}{1.76} femilrat_i + 12.87 fertrate_i, R^2 = 0.7474$$

where *inc* is income per capita, *femiltrat* is the female illiteracy rate, and *fertrate* is the fertility rate

(The numbers in parentheses are standard errors of the estimators.)

a) Test the joint significance of income, illiteracy and fertility rates.

b) Test the significance of the fertility rate.

c) Which of the two models would you choose? Explain your answer.

Exercise 4.19 Using a sample of 32 annual observations, the following estimations were obtained to explain the car sales (*car*) of a particular brand:

$$\widehat{car_i} = \underset{(6.48)}{104.8} - \underset{(3.19)}{6.64} pcar_i + \underset{(0.16)}{2.98} adv_i$$

$$\sum \hat{u}_i^2 = 1805.2; \qquad \sum (car_i - \overline{car})^2 = 13581.4$$

where *pcar* is the price of cars and *adv* are spending on advertising.

(The numbers in parentheses are standard errors of the estimators.)

a) Are price and advertising expenditures significant together? Explain your answer.

b) Can you accept that prices have a negative influence on sales? Explain your answer.

c) Describe in detail how you would test the hypothesis that the impact of advertising expenditures on sales is greater than minus 0.4 times the impact of the price.

Exercise 4.20 In a study of the production costs (*cost*) of 62 coal mines, the following results are obtained:

$$\widehat{cost_i} = \underset{(3.4)}{2.20} - \underset{(0.005)}{0.104}\, dmec_i + \underset{(2.2)}{3.48}\, geodif_i + \underset{(0.15)}{0.104}\, absent_i$$

$$\sum\left[cp_i - \overline{cp}\right]^2 = 109.6 \qquad \sum \hat{u}_i^2 = 18.48$$

where *dmec* is the degree of mechanization, *geodif* is a measurement of geological difficulties and *absent* is the percentage of absenteeism.

a) Test the significance of each of the model coefficients.

b) Test the overall significance of the model.

Exercise 4.21 With fifteen observations, the following estimation was obtained:

$$\hat{y}_i = 8.04 - \underset{(1.00)}{2.46}\, x_{i2} + \underset{(0.60)}{0.23}\, x_{i3}$$

$$\bar{R}^2 = 0.30$$

where the values between parentheses are standard deviations and the coefficient of determination is the adjusted one.

a) Is the coefficient of the variable x_2 significant?

b) Is the coefficient of the variable x_3 significant?

c) Discuss the joint significance of the model.

Exercise 4.22 Consider the following econometric specification:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

With a sample of 26 observations, the following estimations were obtained:

1) $\hat{y}_i = 2 + \underset{(1.9)}{3.5}\, x_{1i} - \underset{(2.2)}{0.7}\, x_{2i} - \underset{(1.5)}{2}\, x_{3i} + u_i$ $\qquad R^2$=0.982

2) $\hat{y}_i = 1.5 + \underset{(2.7)}{3}\, (x_{1i} + x_{2i}) - \underset{(2.4)}{0.6}\, x_{3i} + u_i$ $\qquad R^2$= 0.876

(The *t* statistics are between brackets)

a) Show that the following expressions for the *F*-statistic are equivalent:

$$F = \frac{\left(RSS_R - RSS_{UR}\right)/r}{RSS_{UR}/(n-k)} \qquad F = \frac{\left(R_{UR}^2 - R_R^2\right)/q}{(1-R_{UR}^2)/(n-k)}$$

b) Test the null hypothesis $\beta_2 = \beta_3$.

Exercise 4.23 In the estimation of the Brown model in exercise 3.19, using the workfile *consumsp*, we obtained the following results:

$$\widehat{conspc_t} = -\underset{(84.88)}{7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

$$R^2{=}0.997 \quad RSS{=}1891320 \quad n{=}56$$

Two additional estimations are now obtained:

$$\widehat{conspc_t - conspc_{t-1}} = \underset{(84.43)}{-98.13} + \underset{(0.0803)}{0.2757}(incpc_t - conspc_{t-1})$$

$$R^2{=}0.1792 \quad RSS{=}2199474 \quad n{=}56$$

$$\widehat{conspc_t - incpc_{t-1}} = \underset{(84.88)}{-7.156} - \underset{(0.0090)}{0.0264}\,incpc + \underset{(0.0903)}{0.5771}(conspc_{t-1} - incpc_t)$$

$$R^2{=}0.6570 \quad RSS{=}1891320 \quad n{=}56$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the significance of each of the coefficients for the first model.

b) Test that the coefficient on *incpc* in the first model is smaller than 0.5.

c) Test the overall significance of the first model.

d) Is it admissible that $\beta_2 + \beta_3 = 1$?

e) Show that by operating in the third model you can reach the same coefficients as in the first model.

Exercise 4.24 The following model was formulated to analyze the determinants of the median base salary in $ for graduating classes of 2010 from the best American business schools (*salMBAgr*):

$$salMBAgr = \beta_1 + \beta_2 tuition + \beta_3 salMBApr + u$$

where *tuition* is tuition fees including all required fees for the entire program (but excluding living expenses) and *salMBApr* is the median annual salary in $ for incoming classes in 2010.

Using the data in *MBAtui10*, the previous model has been estimated:

$$\widehat{salMBAgr_i} = \underset{(5415)}{42489} + \underset{(0.0628)}{0.1881}\,tuition_i + \underset{(0.1015)}{0.5992}\,salMBApr_i$$

$$R^2{=}0.703 \quad n{=}39$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the above model are individually significant at 1% and at 5%?

b) Test the overall significance of the model.

c) What is the predicted value of *salMBAgr* for a graduate student who paid 100000$ *tuition* fees in a two-year MBA master and previously had a *salMBApr* equal to 70000$? How many years of work does the student require to offset tuition expenses? To answer this question, suppose that the discount rate equals the expected rate of salary increase and that the student received no wage income during the two master courses.

d) If we added the regressor *rank2010* (the rank of each business school in 2010), the following results were obtained:

$$\widehat{salMBAgr_i} = \underset{(8520)}{61320} + \underset{(0.0626)}{0.1229}\,tuition_i + \underset{(0.1055)}{0.4662}\,salMBApr_i$$

$$\underset{(85.13)}{-232.06}\,rank2010_i$$

$$R^2{=}0.755 \quad n{=}39$$

Which of the regressors included in this model are individually significant at 5%?

What is the interpretation of the coefficient on *rank2010*?

e) The variable *rank2010* is based on three components: *gradpoll* is a rank based on surveys of MBA grads and contributes 45 percent to final ranking; *corppoll* is a rank based on surveys of MBA recruiters and contributes 45 percent to final ranking; and *intellec* is a rank based on a review of faculty research published over a five-year period in 20 top academic journals and faculty books reviewed in *The New York Times*, *The Wall Street Journal*, and *Bloomberg Businessweek* over the same period; this last rank contributes 10 percent to the final ranking. In the following estimated model *rank2010* has been substituted for its three components:

$$\widehat{salMBAgr_i} = \underset{(10700)}{79904} + \underset{(0.0696)}{0.0305}\,tuition_i + \underset{(0.107)}{0.3751}\,salMBApr_i$$

$$-\underset{(94.54)}{303.82}\,gradpoll_i - \underset{(61.26)}{33.829}\,corppoll_i - \underset{(64.09)}{113.36}\,intellec_i$$

$$R^2{=}0.797 \qquad n{=}39$$

What is the weight in percentage of each one of these three components in determining the *salMBAgr*? Compare the results with the contribution of each in defining *rank2010*.

f) Are *gradpoll*, *corppoll* and *intellec* jointly significant at 5%? Are they individually significant at 5%?

Exercise 4.25 (Continuation of exercise 3.12). The population model corresponding to this exercise is:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 tenure + \beta_4 age + u$$

Using workfile *wage06sp*, the previous model was estimated:

$$\widehat{\ln(wage)_i} = \underset{(0.073)}{1.565} + \underset{(0.0035)}{0.0448}\,educ_i + \underset{(0.0019)}{0.0177}\,tenure_i + \underset{(0.0016)}{0.0065}\,age_i$$

$$R^2{=}0.337 \qquad n{=}800$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the overall significance of the model.

b) Is *tenure* statistically significant at 10%? Is *age* positively significant at 10%?

c) Is it admissible that the coefficient of *educ* is equal to that of *tenure*? Is it admissible that the coefficient of *educ* is triple to that of *tenure*? To answer these questions you have the following additional information:

$$\widehat{\ln(wage)_i} = \underset{(0.073)}{1.565} + \underset{(0.0042)}{0.0271}educ_i + \underset{(0.0019)}{0.0177}(educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

$$\widehat{\ln(wage)_i} = \underset{(0.073)}{1.565} - \underset{(0.0071)}{0.0082}\,educ_i + \underset{(0.0019)}{0.0177}(3 \times educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

Can you calculate the R^2 in the two equations in part *c*)? Please do it.

Exercise 4.26 (Continuation of exercise 3.13). Let us take the population model of this exercise as the reference model. In the estimated model, using workfile *housecan*, the standard errors of the coefficients appear between brackets:

$$\widehat{price_i} = -\underset{(3379)}{2418} + \underset{(1207)}{5827}\,bedrooms_i + \underset{(1785)}{19750}\,bathrms_i + \underset{(0.388)}{5.411}\,lotsize_i$$

$$R^2{=}0.486 \qquad n{=}546$$

a) Test the overall significance of this model.

b) Test the null hypothesis that an additional bathroom has the same influence on housing prices than four additional bedrooms. Alternatively, test that an additional bathroom has more influence on housing prices than four additional bedrooms. (Additional information: $\text{var}(\hat{\beta}_2)$ =1455813; $\text{var}(\hat{\beta}_3)$=3186523; and $\text{var}(\hat{\beta}_2,\hat{\beta}_3)$=-764846).

c) If we add the regressor *stories* (number of stories excluding the basement) to the model, the following results have been obtained:

$$\widehat{price_i} = \underset{(3603)}{-4010}+\underset{(1215)}{2825}\,bedrooms_i+\underset{(1734)}{17105}\,bathrms_i$$

$$+\underset{(0.369)}{5.429}\,lotsize_i+\underset{(1008)}{7635}\,stories_i$$

$$R^2\text{=0.536} \quad n\text{=546}$$

What do you think about the sign and magnitude of the coefficient on *stories*? Do you find it surprising? What is the interpretation of this coefficient? Test whether the number of stories has a significant influence on housing prices.

d) Repeat the tests in part *b*) with the model estimated in part *c*). (Additional information: $\text{var}(\hat{\beta}_2)$=1475758; $\text{var}(\hat{\beta}_3)$=3008262; and $\text{var}(\hat{\beta}_2,\hat{\beta}_3)$=-554381).

Exercise 4.27 (Continuation of exercise 3.14). Let us take the population model of this exercise as the reference model. Using workfile *ceoforbes*, the estimated model was the following:

$$\widehat{\ln(salary)_i} = \underset{(0.377)}{4.641}+\underset{(0.0033)}{0.0054}\,roa_i+\underset{(0.0425)}{0.2893}\ln(sales_i)+\underset{(0.0000220)}{0.0000564}\,profits_i+\underset{(0.0032)}{0.0122}\,tenure_i$$

$$R^2\text{=0.232} \quad n\text{=447}$$

(The numbers in parentheses are standard errors of the estimators.)

a) Does *roa* have a significant effect on salary? Does *roa* have a significant positive effect on salary? Carry out both tests at the 10% and 5% significance level.

b) If *roa* increases by 20 points, by what percentage is *salary* predicted to increase?

c) Test the null hypothesis that the elasticity *salary/sales* is equal to 0.4.

d) If we add the regressor *age*, the following results are obtained:

$$\widehat{\ln(salary)_i} = \underset{(0.442)}{4.159}+\underset{(0.0033)}{0.0055}\,roa_i+\underset{(0.0423)}{0.2903}\ln(sales_i)+\underset{(0.0000220)}{0.0000539}\,profits$$

$$+\underset{(0.0035)}{0.00924}\,tenure_i+\underset{(0.0043)}{0.00880}\,age_i$$

$$R^2\text{=0.240} \quad n\text{=447}$$

Are the estimated coefficients very different from the estimates in the reference model? What about the coefficient on *tenure*? Explain it.

e) Does *age* have a significant effect on the salary of a CEO?

f) Is it admissible that the coefficient of *age* is equal to the coefficient of *tenure*? (Additional information: $\text{var}(\hat{\beta}_5)$=1.24E-05; $\text{var}(\hat{\beta}_6)$=1.82E-05; and $\text{var}(\hat{\beta}_5,\hat{\beta}_6)$=-6.09E-06).

Exercise 4.28 (Continuation of exercise 3.15). Let us take the population model of this exercise as the reference model. Using workfile *rdspain*, the estimated model was the following:

$$\widehat{rdintens_i} = \underset{(0.428)}{-1.8168} + \underset{(0.0278)}{0.1482}\ln(sales_i) + \underset{(0.0021)}{0.0110}\,exponsal_i$$

$$R^2\text{=}0.048 \qquad n\text{=}1983$$

(The numbers in parentheses are standard errors of the estimators.)

a) Is the *sales* variable individually significant at 1%?

b) Test the null hypothesis that the coefficient on *sales* is equal to 0.2?

c) Test the overall significance of the reference model.

d) If we add the regressor ln(*workers*), the following results are obtained:

$$\widehat{rdintens} = \underset{(0.750)}{0.480} - \underset{(0.0687)}{0.08585}\ln(sales) + \underset{(0.0021)}{0.01049}\,exponsal + \underset{(0.09198)}{0.3422}\ln(workers)$$

$$R^2\text{=}0.055 \qquad n\text{=}1983$$

Is *sales* individually significant at 1% in the new estimated model?

e) Test the null hypothesis that the coefficient on ln(*workers*) is greater than 0.5?

Exercise 4.29 (Continuation of exercise 3.16). Let us take the population model of this exercise as the reference model. Using workfile *hedcarsp*, the corresponding fitted model is the following:

$$\widehat{\ln(price)_i} = \underset{(0.154)}{14.42} + \underset{(0.0000438)}{0.000581}\,cid_i + \underset{(0.0079)}{0.003823}\,hpweight_i - \underset{(0.0122)}{0.07854}\,fueleff_i$$

$$R^2\text{=}0.830 \qquad n\text{=}214$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 1%?

b) Add the variable *volume* to the reference model. Does *volume* have a statistically significant effect on ln(*price*)? Does *volume* have a statistically significant positive effect on ln(*price*)?

c) Is it admissible that the coefficient of *volume* estimated in part *b*) is equal but is the opposite of the coefficient of *fueloff*?

d) Add the variables *length*, *width* and *height* to the model estimated in part *b*). Taking into account that *volume*=*length*×*width*×*height*, is there perfect multicollinearity in the new model? Why? Why not? Estimate the new model if it is possible.

e) Add the variable ln(*volume*) to the reference model. Test the null hypothesis that the *price*/*volume* elasticity is equal to 1?

f) What happens if you add the regressors ln(*length*), ln(*width*) and ln(*height*) to the model estimated in part *e*)?

Exercise 4.30 (Continuation of exercise 3.17). Let us take the population model of this exercise as the reference model. Using workfile *timuse03*, the corresponding fitted model is the following:

$$\widehat{houswork_i} = \underset{(23.27)}{141.9} + \underset{(1.621)}{3.850}\,educ_i - \underset{(0.00539)}{0.00917}\,hhinc_i + \underset{(0.311)}{1.767}\,age_i - \underset{(0.0229)}{0.2289}\,paidwork_i$$

$$R^2\text{=}0.1440 \qquad n\text{=}1000$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 5% and at 1%?

b) Estimate a model in which you could test directly whether one additional year of education has the same effect on time devoted to house work as two additional years of age. What is your conclusion?

c) Test the joint significance of *educ* and *hhnc*.

d) Run a regression in which you add the variable *childup3* (number of children up to three years) to the reference model. In the new model, which of the regressors are individually significant at 5% and at 1%?

e) In the model formulated in *d*), what is the most influential variable? Why?

Exercise 4.31 (Continuation of exercise 3.18). Let us take the population model of this exercise as the reference model. Using workfile *hdr2010*, the corresponding fitted model is the following:

$$\widehat{stsfglo_i} = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\,gnipc_i + \underset{(0.009)}{0.0858}\,lifexpec_i$$

$$R^2=0.642 \quad n=144$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 1%?

b) Run a regression by adding the variables *popnosan* (population in percentage without access to improved sanitation services) and *gnirank* (rank in *gni*) to the reference model. Which of the regressors included in the new model are individually significant at 1%? Interpret the coefficients on *popnosan* and *gnirank*.

c) Are *popnosan* and *gnirank* jointly significant?

d) Test the overall significance of the model formulated in *b*).

Exercise 4.32 Using a sample of 42 observations, the following model has been estimated:

$$\hat{y}_t = -670.591 + 1.008x_t$$

For observation 43, it is known that the value of *x* is 1571.9.

a) Calculate the point predictor for observation 43.

b) Knowing that the variance of the prediction error $\hat{e}_2^{43} = y^{43} - \hat{y}^{43}$ is equal to $(24.9048)^2$, calculate a 90% probability interval for the individual value.

Exercise 4.33 Besides the estimation presented in exercise 4.23, the following estimation on the Brown consumption function is also available:

$$\widehat{conspc_t} = \underset{(64.35)}{12729} + \underset{(0.0857)}{0.3965}(incpc_t - 13500) + \underset{(0.0903)}{0.5771}(conspc_{t-1} - 12793.6)$$

$$R^2=0.997 \quad RSS=1891320 \quad n=56$$

(The numbers in parentheses are standard errors of the estimators.)

a) Obtain the point predictor for consumption per capita in 2011, knowing that $incpc_{2011}$=13500 and $conspc_{2010}$=12793.6.

b) Obtain a 95%confidence interval for the expected value of consumption per capita in 2011.

c) Obtain a 95% prediction interval for the individual value of consumption per capita in 2011.

Exercise 4.34 (Continuation of exercise 4.30) Answer the following questions:

a) Using the first estimation in exercise 4.30, obtain a prediction for *houswork* (minutes devoted to house-work per day), when you plug in the equation *educ*=10 (years), *hhinc*=1200 (euros per month), *age*=50 (years) and *paidwork*=400 (minutes per day).

b) Run a regression, using workfile *timuse03*, which allows you to calculate a 95% CI with the characteristics used in part *a)*.

c) Obtain a 95% prediction interval for the individual value of *houswork* with the characteristics used in parts *a)*.

Exercise 4.35 (Continuation of exercise 4.29) Answer the following questions:

a) Plug in the first equation of the exercise 4.29 of *cid*=2000 (cubic inch displacement), *hpweight*=10 (ratio horsepower/weight in kg expressed as percentage), and *fueleff*=6 (minutes per day) Obtain the point predictor of consumption per capita in 2011, knowing that $incpc_{2011}$=12793.6 and $conspc_{2010}$=13500.

b) Obtain a consistent estimate of *price* with the characteristics used in parts *a)*.

c) Run a regression that allows you to calculate a 95% CI with the characteristics used in part a).

d) Obtain a 95% prediction interval for the individual value of the consumption per capita 2011.