6 Relaxing the assumptions in the linear classical model

Ezequiel Uriel
University of Valencia
Version: 09-2013

6.1 Relaxing the assumptions in the linear classical model: an overview

In chapters 2 and 3, single and multiple linear regression models were formulated, including the set of statistical assumptions called the classical linear model (*CLM*) assumptions. Now, let us examine the problems posed by the failure of each one of the *CLM* assumptions and alternative methods for estimating the linear model.

Assumption on the functional form

Assumption 1 postulates the following population model:

$$y=\beta_1+\beta_2 x_1+\cdots+\beta_k x_k+u \qquad (6\text{-}1)$$

This assumption specifies what the endogenous variable is and its functional form, as well as what the explanatory variables are and their functional forms. It also states that the model is linear on the parameters

If we estimate a different population model, a misspecification error is made. The consequences of such errors will be discussed in section 6.2.

Assumptions on the regressors

The assumptions 2, 3 and 4 were made on the regressors. In the multiple linear regression, assumption 2 postulated that the values $x_2, x_3, \cdots, x_k$ are fixed in repeated samples, that is to say, the regressors are non-stochastic. This is a reasonable assumption when the regressors are obtained from experiments. However, it is less admissible for variables obtained by observation in a passive way, as in the case of income in the consumption function.

When the regressors are stochastic, the statistical relationship between the regressors and the random disturbance is crucial in building an econometric model. For this reason, an alternative assumption was formulated as 2*: the regressors $x_2, x_3, \cdots, x_k$ are distributed independently of the random disturbance. When we assume this alternative assumption, the inference, conditional on the matrix of regressors, leads to results that are virtually coincident with the case where the matrix **X** is fixed. In other words, in the case of independence between the regressors and the random disturbance, the ordinary least squares method is still the optimal method for estimating the vector of coefficients.

In assumption 3 it was postulated that the matrix of regressors **X** contains no measurement errors. If there are measurement errors, a very serious econometric problem will arise with a complex solution.

Assumption 4 states that there is no exact linear relationship between the regressors, or, in other words, it establishes that there is no perfect multicollinearity in the model. This assumption is necessary to calculate the *OLS* estimators. Perfect multicollinearity is not used in practice. Instead, there is often an approximately linear relationship between the regressors. In this case the estimators obtained will not be accurate, although they still retain the property of being *BLUE* estimators. In other words, the relationship between the regressors makes it difficult to quantify the effect that each one has on the regressand. This is due to the fact that the variances of the estimators are high. When an approximately linear relationship between the regressors exists, multicollinearity is not perfect. Section 6.3 will be devoted to examining the detection of non-perfect multicollinearity, along with some possible solutions

Assumptions on the parameters

In assumption 5 it was assumed that the parameters are not random. The real world suggests that this coefficient constancy is not reasonable. In models using time series data, there are often changes in patterns of behavior over time, which would naturally involve changes in the regression coefficients. In any case, section 5.6 examines the test of structural change which determines whether there has been any change in the parameters over time.

Assumptions on the random disturbance term

In assumption 6 it is assumed that $E(\mathbf{u})=\mathbf{0}$. This assumption is not empirically testable in the general case of models with intercept.

Before moving on to other assumptions on the random disturbance u_i, it should be noted that this is an unobservable variable. Information on u_i is obtained indirectly through the residuals, which will be used for testing the behavior of the disturbances. However, the use of residuals to perform tests on disturbances poses some problems. When the *CLM* assumptions are fulfilled, the random disturbances are neither autocorrelated nor homoskedastic, whereas the residuals are heteroskedastic and autocorrelated under these assumptions. These circumstances are important in the design of statistical tests on heteroskedasticity and no autocorrelation.

If assumptions 7 of homoscedasticity and/or 8 of no autocorrelation are not fulfilled, the least squares estimators are still linear and unbiased but they are not the best.

The assumptions of homoskedasticity and no autocorrelation formulated in chapter 3, respectively, may be formulated together indicating that the covariance matrix of random disturbances is a scalar matrix, i.e.:

$$E(\mathbf{uu}') = \sigma^2\mathbf{I} \tag{6-2}$$

When one or both assumptions indicated are not fulfilled, then the covariance matrix will be less restrictive. Thus, we will consider the following covariance matrix of the disturbances:

$$E(\mathbf{uu}') = \sigma^2\boldsymbol{\Omega} \tag{6-3}$$

where the only restriction imposed on $\boldsymbol{\Omega}$ is that it is a positive definite matrix

When the covariance matrix is a non-scalar matrix such as (6-3), then one can obtain linear, unbiased and best estimators by applying the method of generalized least squares (*GLS*). The expression of these estimators is as follows:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} \tag{6-4}$$

In practice, formula (6-4) is not directly applied. Instead a two-step process that leads to exactly the same results is applied.

In section 6.5, we will examine the tests to determine whether there is heteroskedasticity, as well as the particularization of the *GLS* method in this case. Section 6.6 will present testing methods and the appropriate treatment of autocorrelation.

Assumption 9 of normality postulated in the *CLM* allows us to make statistical inferences with known distributions. If the normality assumption is not adequate, then the tests will only be approximately valid. In section 6.4, a normality test of the disturbances is used to determine whether this assumption is acceptable.

6.2 Misspecification

Misspecification occurs when we estimate a different model from the population model. The problem in social sciences, and in particular in economics, is that we do not usually know the population model.

Bearing in mind this observation, we shall consider three types of misspecification:

- Inclusion of irrelevant variables.
- Exclusion of relevant variables.
- Incorrect functional form.

6.2.1 Consequences of misspecification

We will examine the consequences of each type of misspecification on the *OLS* estimators

Inclusion of an irrelevant variable

Let us consider, for example, that the population model is the following:

$$y = \beta_1 + \beta_2 x_2 + u \tag{6-5}$$

Consequently, the *population regression function* (*PRF*) is given by

$$\mu_y = \beta_1 + \beta_2 x_2 \tag{6-6}$$

Now let us suppose that the *sample regression function* (*SRF*) estimated is the following

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} + \tilde{\beta}_3 x_{3i} \tag{6-7}$$

This is the case of *inclusion of an irrelevant variable*: specifically, in (6-7) we have introduced the irrelevant variable x_3. What are the effects of including an irrelevant variable in the *OLS* estimators?

It can be shown that the estimators corresponding to (6-7) are unbiased, that is to say,

$$E(\tilde{\beta}_1) = \beta_1 \qquad E(\tilde{\beta}_2) = \beta_2 \qquad E(\tilde{\beta}_3) = 0$$

However, the variances of these estimators will be greater than those obtained by estimating (6-5) in which x_3 is (correctly) omitted.

This result can be extended to the case of including one or more irrelevant variables. In this case *OLS* estimators are unbiased, but with variances greater than when the irrelevant variables are not included in the estimated model.

Exclusion of a relevant variable

Let us consider, for example, that the population model is the following:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \tag{6-8}$$

The *PRF* is therefore given by:

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6-9}$$

Now let us suppose that the *SRF* we estimate, due to ignorance or data unavailability, is the following

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \tag{6-10}$$

This is a case of *exclusion of a relevant variable*: in (6-10) we have omitted the relevant variable x_3. Is $\tilde{\beta}_2$, obtained by applying *OLS* in (6-10), an unbiased estimator of β_2?

As appendix 6.1 shows, the estimator $\tilde{\beta}_2$ is biased. The bias is

$$Bias(\tilde{\beta}_2) = \beta_3 \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2} \tag{6-11}$$

The bias is null if, according to (6-11), the covariance between x_2 and x_3 is 0. It is important to remark that the ratio

$$\frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}$$

is just the *OLS* slope ($\hat{\delta}_2$) coefficient from regression of x_3 on $x_{2.}$ That is to say,

$$\hat{x}_2=\hat{\delta}_1+\hat{\delta}_2\hat{x}_2=\hat{\delta}_1+\frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}\hat{x}_2 \qquad (6\text{-}12)$$

Thus, according to (6-72) - in appendix 6.1-, and (6-12), we can write that

$$E(\tilde{\beta}_2)=\beta_2+\beta_3\hat{\delta}_2 \qquad (6\text{-}13)$$

Therefore, the bias is equal to $\beta_3\hat{\delta}_2$. In table 6.1, there is a summary of the sign of the bias in $\tilde{\beta}_2$ when x_3 is omitted in estimating equation. It must be taken into account that the sign of $\hat{\delta}_2$ is the same as the sign of the sample correlation between x_2 and x_3.

TABLE 6.1. Summary of bias in $\tilde{\beta}_2$ when x_3 is omitted in estimating equation.

	Corr(x_2,x_3)>0	*Corr(x_2,x_3)<0*
$\beta_3>0$	Positive bias	Negative bias
$\beta_3<0$	Negative bias	Positive bias

Incorrect functional form

If we use a functional form different from the true population model, then the *OLS* estimators will be biased.

In conclusion, if there is exclusion of relevant variables or/and an incorrect functional form has been used, then the *OLS* estimators will be biased and also inconsistent. Therefore, the conventional inference procedures will be invalidated in these two cases.

6.2.2 Specification tests: the RESET test

To test whether irrelevant variables are included in the model we can apply the exclusion restriction tests, which we have examined in chapter 4.

To test the exclusion of relevant variables or the use of an incorrect functional form, we can apply the RESET (Regression Equation Specification Error Test) test. This test is a general test for specification errors proposed by Ramsey (1969). In order to explain it, consider that the *initial* model is the following:

$$y=\beta_1+\beta_2x_2+\beta_3x_3+u \qquad (6\text{-}14)$$

Now, we introduce an *augmented* model in which two new variables (z_1 and z_2) appear:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 z_1 + \alpha_2 z_2 + u \tag{6-15}$$

Taking into account the specification of the two models, the null and alternative hypotheses will be the following:

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = 0 \\ &H_1 : H_0 \text{ is not true} \end{aligned} \tag{6-16}$$

The crucial question in building the test is to determine the *z* variables or regressors to be introduced. In the case of exclusion of relevant variables, the *z* variables will be the omitted regressors which may be new variables or also squares and powers of previous variables. The test to be applied would be similar to the exclusion tests, but with the roles reversed: the restricted model is now the *initial* model, while the unrestricted model corresponds to the *augmented* model.

In testing for incorrect functional form, consider, for example, that (6-14) is specified instead of the true relationship:

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + u \tag{6-17}$$

In model (6-17), there is a multiplicative relationship between the regressors. Ramsey took into account that a Taylor series approximation of the multiplicative relationship would yield an expression involving powers and cross-products of the explanatory variables. For this reason, he suggests including, in the augmented model, powers of the predicted values of the dependent variable (which are, of course, linear combinations of power and cross-product terms of the explanatory variables):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 \hat{y}^2 + \alpha_2 \hat{y}^3 + u \tag{6-18}$$

where the $\hat{y}$´s are the *OLS* fitted values corresponding to the model (6-14). The superscripts indicate the powers to which these predictions are raised. The first power is not included since it is perfectly collinear with the rest of the regressors of the initial model.

The steps involved in the RESET test are as follows:

Step 1. The *initial* model is estimated and the *fitted values,* $\hat{y}_i$, are calculated.

Step 2. The *augmented* model, which can include one or more powers of $\hat{y}_i$, is estimated.

Step 3. Taking the R^2_{init} corresponding to the initial model and the R^2_{augm} corresponding to the augmented model, the *F* statistic is calculated:

$$F = \frac{(R^2_{augm} - R^2_{init})/r}{(1 - R^2_{augm})/(n-h)} \tag{6-19}$$

where *r* is the number of new parameters added to the initial model, and *h* is the number of parameters of the augmented model, including the intercept.

Under the null hypothesis, this statistic is distributed as follows:

$$F \mid H_0 \sim F_{r,n-h} \tag{6-20}$$

Step 4. For a significance level α, and designating by $F^{\alpha}_{r,n-h}$ the corresponding value in the F table, the decision to make is the following:

$$\begin{array}{llll} \text{If} & F \geq F^{\alpha}_{r,n-h} & \text{reject} & H_0 \\ \text{If} & F < F^{\alpha}_{r,n-h} & \text{not reject} & H_0 \end{array}$$

Therefore, high values of the statistic lead to the rejection of the initial model.

In RESET test we test the null hypothesis against an alternative hypothesis that does not indicate what the correct specification should be. This test is therefore a misspecification test which may indicate that there is some form of misspecification but does not give any indication of what the correct specification should be.

EXAMPLE 6.1 Misspecification in a model for determination of wages

Using a subsample of data from the *wage structure survey* of Spain for 2006 (file *wage06sp*), the following model is estimated:

$$\widehat{wage}_i = \underset{(1.55)}{4.679} + \underset{(0.146)}{0.681} educ_i + \underset{(0.071)}{0.293} tenure_i$$

$$R^2 = 0.249 \quad n = 150$$

where *educ* (education) and *tenure* (experience in the firm) are measured in years and *wage* in euros per hour.

Considering that we may have a problem of incorrect functional form, an augmented model is estimated. In this augmented model - besides *educ*, *tenure*, and the intercept - $\widehat{wage}_i^2$ and $\widehat{wage}_i^3$ from the initial model are included as regressors. The F statistic calculated using the R^2_{init} and R^2_{augm}, according to (6-19), is equal to 4.18. Given that $F^{0.05}_{2,145} \simeq F^{0.05}_{2,60} = 3.15$, we reject that, for the levels α=0.05 and α=0.10, the linear form is adequate to explain wage determination. On the contrary, given that $F^{0.01}_{2,145} \simeq F^{0.01}_{2,60} = 4.98$ H_0 is not rejected for α=0.01.

6.3 Multicollinearity

6.3.1 Introduction

Perfect multicollinearity is not usually seen in practice, unless the model is wrongly designed as we saw in chapter 5. Instead, an approximately linear relationship between the regressors often exists. In this case, the estimators obtained will generally not be very accurate, despite still being *BLUE*. In other words, the relationship between regressors makes it difficult to quantify accurately the effect each one has on the regressand. This is due to the fact that the variances of the estimators are high. When there is an approximately linear relationship between the regressors, then it is said that there is *not perfect multicollinearity*. The multicollinearity problem arises because there is insufficient information to get an accurate estimation of model parameters.

To analyze the problem of multicollinearity, we will examine the variance of an estimator. In the multiple linear regression model, the estimator of the variance of any slope coefficient - for example, $\hat{\beta}_j$ - is equal, as we saw in (3-68), to

$$\widehat{\text{var}(\hat{\beta}_j)} = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{6-21}$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2, n is the sample size, S_j^2 is the sample variance of the regressor x_j, and R_j^2 is the R-squared obtained from regressing x_j on all other x's.

The last of these four factors which determines the value of the variance of $\hat{\beta}_j$, (1-R_j^2), is precisely an indicator of multicollinearity. Multicollinearity arises in estimating β_j when R_j^2 is "close" to one, but there is no absolute number that we can quote to conclude that multicollinearity is really a problem for the precision of the estimators. Although the problem of multicollinearity cannot be clearly defined, it is true that, for estimating β_j, the lower the correlation between x_j and the other independent variables the better. If R_j^2 is equal to 1, then we would have perfect multicollinearity and it is not possible to obtain the estimators of the coefficients. In any case, when one or more R_j^2 are close to 1, multicollinearity is a serious problem. In this case, when making inferences with the model, the following problems arise:

a) The variances of the estimators are very large.

b) The estimated coefficients will be very sensitive to small changes in the data.

6.3.2 Detection

Multicollinearity is a problem of the *sample*, because it is associated with the specific configuration of the sample of the x's. For this reason, there are no statistical tests. (Remember that statistical tests only work with *population* parameters). Instead, many practical rules were developed attempting to determine to what extent multicollinearity seriously affects the inference made with a model. These rules are not always reliable, and in some cases are questionable. In any case, we are going to look at some measures that are very useful to detect the degree of multicollinearity: the *variance inflation factor* (*VIF*) and the *tolerance*, and the *condition number* and the *coefficient variance decomposition*.

Variance inflation factor (VIF) and tolerance

In order to explain the meaning of these measures, let us suppose there is *no* linear relationship between x_j and the other explanatory variables in the model, that is to say, the regressor x_j is *orthogonal* to the remaining regressors. In this case, R_j^2 will be zero and the variance of $\hat{\beta}_j$ will be

$$\widehat{\text{var}(\beta_j^*)} = \frac{\hat{\sigma}^2}{nS_j^2} \tag{6-22}$$

Dividing (6-21) by (6-22), we obtain the variance inflation factor (*VIF*) as

$$VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2} \tag{6-23}$$

The *VIF* statistic calculated according to (6-23) is sometimes called "centered *VIF*" to be distinguished from the "uncentered *VIF*" which is interesting in models without intercept. The E-views programme supplies both statistics.

Tolerance, which is the inverse of *VIF*, is defined as

$$Tolerance(\hat{\beta}_j) = \frac{1}{VIF} = 1 - R_j^2 \tag{6-24}$$

Thus, $VIF(\hat{\beta}_j)$ is the ratio between the estimated variance and the one that there would have been if x_j was uncorrelated with the other regressors in the model. In other words, the *VIF* shows the extent to which the variance of the estimator is "inflated" as a result of non-orthogonallity of the regressors. It is readily seen that the higher the *VIF* (or the lower the tolerance index), the higher the variance of $\hat{\beta}_j$.

The procedure is to choose each one of the regressors at a time as the dependent variable and to regress them against a constant and the remaining explanatory variables. We would then get *k* values for the *VIF*'s. If any of them is high, then multicollinearity is detected. Unfortunately, however, there is no theoretical indicator to determine whether the *VIF* is "high." Also, there is no theory that tells us what to do if multicollinearity is found.

The variance inflation factor (*VIF*) and the tolerance are both widely used measures of the degree of multicollinearity. Unfortunately, several rules of thumb – most commonly the rule of 10 – associated with the *VIF*– are regarded by many practitioners as a sign of severe or serious multicollinearity (this rule appears in both scholarly articles and advanced statistical textbooks), but this rule has no scientific justification

The problem with the *VIF* (or the tolerance) is that it does not provide any information that could be used to treat the problem.

EXAMPLE 6.2 Analyzing multicollinearity in the case of labor absenteeism

In example 3.1 a model was formulated and estimated, using file *absent*, to explain absenteeism from work as a function of the variables *age, tenure* and *wage*.

Table 6.2 provides information on the tolerance and the *VIF* of each regressor. According to these statistics, multicollinearity does not appear to affect the *wage* but there is a certain degree of multicollinearity in the variables *age* and *tenure*. In any case, the problem of multicollinearity in this model does not appear to be serious because all *VIF* are below 5.

TABLE 6.2. Tolerance and *VIF*.

	Collinearity statistics	
	Tolerance	*VIF*
age	0.2346	4.2634
tenure	0.2104	4.7532
wage	0.7891	1.2673

Condition number and coefficient variance decomposition

This method, developed by Belsey *et al.* (1982), is based on the variance decomposition of each regression coefficient as a function of the eigenvalues λ_h of the matrix **X'X** and the corresponding elements of the associate eigenvectors. We will not discuss eigenvalues and eigenvectors here, because they are beyond the scope of this book, but in any case we will see their application.

The *condition number* is a standard measure of ill-conditioning in a matrix. It indicates the potential sensitivity of the computed inverse matrix to small changes in the original matrix (**X'X** in the case of the regression). Multicollinearity reveals its presence by one or more eigenvalues of **X'X** being "small". The closer a matrix is to singularity the smaller the eigenvalues. The condition number (κ) is defined as the square root of the largest eigenvalue ($\lambda_{\max}$) divided by the smallest eigenvalue ($\lambda_{\min}$):

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \qquad (6\text{-}25)$$

When there is no multicollinearity at all, then all the eigenvalues and the condition number will be equal to one. As multicollinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem), and the condition number will increase. An informal rule of thumb is that if the condition number is greater than 15, multicollinearity is a concern; if it is greater than 30 multicollinearity is a very serious concern.

The variance of $\hat{\beta}_j$ can be decomposed into the contributions from each one of the eigenvalues and can be expressed in the following way:

$$\text{var}(\hat{\beta}_j) = \sigma^2 \sum_h \frac{u_{jh}^2}{\lambda_h} \qquad (6\text{-}26)$$

Thus, the proportion of the contribution of eigenvalue λ_h in the variance of $\hat{\beta}_j$ is equal to

$$\phi_{jh} = \frac{\dfrac{u_{jh}^2}{\lambda_h}}{\displaystyle\sum_{h=0}^{k} \frac{u_{jh}^2}{\lambda_h}} \qquad (6\text{-}27)$$

High values of ϕ_{jh} indicate that, as a consequence of multicollinearity, there is an inflation of the variance. Given that eigenvalues close to zero indicate a multicollinearity problem, it is important to pay special attention to the contribution of the smallest eigenvalues. The contributions corresponding to the smallest eigenvalue may give a clue of the regressors which are involved in the multicollinearity problem.

EXAMPLE 6.3 ***Analyzing the multicollinearity of factors determining time devoted to housework***

In order to analyze the factors that influence time devoted to housework, the following model was formulated in exercise 3.17, using file *timuse03*:

$$houswork = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

where *educ* is the years of education attained, and *hhinc* is the household income in euros per month. The variables *houswork* and *paidwork* are measured in minutes per day.

Table 6.3 provides information on eigenvalues, sorted from the smallest to the largest, and the variance decomposition proportions for each eigenvalue are calculated according to (6-27). The condition number is equal to

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{542.14}{7.06E-06}} = 8782$$

The condition number is very big, which would indicate a large amount of multicollinearity.

As can be seen in table 6.3[1], the greater proportions associated with the smallest eigenvalue, which is the main cause of multicollinearity in this model, correspond to the regressors *educ* and *age*. These two regressors are inversely correlated. The greatest proportions associated with the second smallest eigenvalue correspond to the regressors *educ* and the household income, which are positively correlated.

TABLE 6.3. Eigenvalues and variance decomposition proportions.

Eigenvalues	7.03E-06	0.000498	0.025701	1.861396	542.1400

Variance decomposition proportions

	Associated Eigenvalue				
Variable	1	2	3	4	5
C	0.999995	4.72E-06	8.36E-09	1.23E-13	1.90E-15
EDUC	0.295742	0.704216	4.22E-05	2.32E-09	3.72E-11
HHINC	0.064857	0.385022	0.209016	0.100193	0.240913
AGE	0.651909	0.084285	0.263805	5.85E-07	1.86E-08
PAIDWORK	0.015405	0.031823	0.007178	0.945516	7.80E-05

6.3.3 Solutions

In principle, the problem of multicollinearity is related to deficiencies in the sample. The non-experimental design of the sample is often responsible for these deficiencies. Let us look at some of the solutions to solve the problem of multicollinearity.

Elimination of variables

Multicollinearity can be mitigated if the regressors most affected by multicollinearity are removed. The problem with this solution is that the estimators of the new model would be biased if the original model was correct. On this issue the following reflection should be made. In any case, the researcher is interested in obtaining an unbiased estimator (or at least with very small bias) with a reduced variance. The mean square error (*MSE*) includes both factors. Thus, for the estimator $\hat{\beta}_j$, the *MSE* is defined as follows:

$$MSE(\hat{\beta}_j) = \left[bias(\hat{\beta}_j)\right]^2 + var(\hat{\beta}_j) \qquad (6\text{-}28)$$

If a regressor is eliminated from the model, the estimator of a regressor that is maintained (for example, $\hat{\beta}_j$) will be biased. Nevertheless, its *MSE* can be lower than that of the original model, because the omission of a variable can sufficiently reduce the variance of the estimator. In sum, although the elimination of a variable is not a desirable practice in principle, under certain circumstances it can be justified when it contributes to decreasing the *MSE*.

Increasing the sample size

Given that some degree of multicollinearity is a problem particularly when the variances of the estimators increase significantly, the solutions should aim to reduce

[1] In table 6.3, the eigenvalues are ordered from the lowest to the highest as the associated eigenvalues in the variance decomposition proportions. It is important to remark that in E-views eigenvalues are ordered from the highest to the lowest. However, in this package the condition number is defined differently than usual in the econometrics manuals which we have followed.

these variances. A solution for increasing the variability of the regressors across the sample consists in introducing additional observations. However, this is not always feasible, since the data used in empirical analysis generally come from different data sources given the researcher only collects information on rare occasions.

Furthermore, when dealing with experimental designs, the variability of the regressors can be directly increased without increasing the size of the sample.

Using outside sample information

Another possibility is the use of outside sample information, either by setting constraints on the parameters of the model, or by using estimates from other studies.

Establishing restrictions on the parameters of the model reduces the number of parameters to be estimated and therefore alleviates the possible shortcomings of the sample information. In any case, these restrictions must be inspired by the theoretical model itself, or at least have an economic meaning.

In general, a disadvantage of this approach is that the meaning attributed to the estimator obtained in cross sectional data is very different from that obtained with time series data, in the case when both types of data are jointly used. Sometimes these estimators can be truly "foreign" or outside the object of study.

Using ratios

If instead of the regressand and the regressors of the original model, we use ratios with respect to the most affected regressor by collinearity, the correlations among the regressors of the model may decrease. One such solution is very attractive for the simplicity of implementation. However, the transformations of the original variables of the model using ratios can cause other problems. Assuming the original model fulfills the *CLM* assumptions, this transformation implicitly modifies the properties of the model, and therefore the disturbances of the transformed model will no longer be homoskedastic but heteroskedastic.

6.4 Normality test

The *F* and *t* significance tests built in chapter 4 are based on the normality assumption of the disturbances. But it is not usual to perform a normality test, given that a sufficiently large sample -e.g. 50 or more observations - is not often available. However, normality tests have recently been receiving a growing interest in both theoretical and applied studies.

Let us examine one test for verifying the assumptions of normality of disturbances in an econometric model. This test was proposed by Bera and Jarque, and is based on the statistics of skewness and kurtosis of the residuals.

The skewness statistic is the standardized third-order moment, applied to the residuals, and its expression is the following:

$$\gamma_{1(\hat{u})} = \frac{\sum \hat{u}_i^3 / n}{\left[\sum \hat{u}_i^2 / n\right]^{3/2}} \tag{6-29}$$

In a symmetric distribution, as is the case of the normal distribution, the coefficient of skewness is 0.

The kurtosis statistic is the standardized fourth-order moment, applied to residuals, and its expression is the following:

$$\gamma_{2(\hat{u})} = \frac{\sum \hat{u}_i^4 / n}{\left[\sum \hat{u}_i^2 / n\right]^2} \tag{6-30}$$

In a standard normal distribution, i.e. in an *N*(0.1), the coefficient of kurtosis is equal to 3.

The Bera and Jarque statistic *(BJ*) is given by:

$$BJ = \left[\frac{n}{6}\left(\gamma_{1(\hat{u})}\right)^2 + \frac{n}{24}\left(\gamma_{2(\hat{u})} - 3\right)^2\right] \tag{6-31}$$

In a theoretical normal distribution, the above expression will be equal to 0, as the coefficient of skewness and kurtosis respectively take the values 0 and 3. The statistic *BJ* will take higher values as the coefficient of asymmetry is far from 0 and the coefficient of kurtosis is far from 3. Under the null hypothesis of normality, the statistic *BJ* has the following distribution

$$BJ \xrightarrow[n\to\infty]{} \chi_2^2 \tag{6-32}$$

The indication $n \to \infty$ means that *BJ* is an asymptotic test, i.e. valid when the sample is sufficiently large.

EXAMPLE 6.4 Is the hypothesis of normality acceptable in the model to analyze the efficiency of the Madrid Stock Exchange?

In example 4.5, using file *bolmadef*, we analyzed the market efficiency of the Madrid Stock Exchange in 1992, using a model that relates the daily rate of return on the rate of the previous day. Now we will test the normality assumption on the disturbances of this model. Given the low proportion of the variance explained with this model (see example 4.5), the test of normality of the disturbances is roughly equivalent to test the normality of the endogenous variable.

Table 6.4 shows the coefficients of skewness, kurtosis and the Bera and Jarque statistic, applied to the residuals. The asymmetry coefficient (-0.04) is not far from the value 0 corresponding to a distribution *N*(0.1). On the other hand, the coefficient of kurtosis (4.43) is slightly different from 3, which is the value in the normal distribution. In this case, we reject the assumption of normality for the usual levels of significance, as the Bera and Jarque statistic takes the value of 21.02, which is larger than $\chi_2^{2(0.01)}$ = 9.21.

TABLE 6.4. Normality test in the model on the Madrid Stock Exchange.

skewness coefficient	*kurtosis coefficient*	*Bera and Jarque statistic*
-0.0421	4.4268	21.0232

The fact that the normality assumption is rejected may seem paradoxical, since the values of kurtosis and especially of skewness do not differ substantially from the values taken by these coefficients in a normal distribution. However, the discrepancies are significant enough because they are supported by a large sample size (247 observations). If *n* (the size of the sample) had been 60 rather than 247, the *BJ* statistic, calculated according to (6-31) and using the same coefficient of skewness and kurtosis, takes the value of 5.11, which is smaller than $\chi_2^{2(0.01)}$ = 9.21. To put it another way, with the same coefficients, but with a smaller sample, there is not enough empirical evidence to reject the null hypothesis of normality. Note that this is due to the fact that the *BJ* statistic increases proportionally to the size of the sample, but the degrees of freedom (2) remain unchanged.

6.5 Heteroskedasticity

The homoskedasticity assumption (assumption 7 of the *CLM*) states that the disturbances have a constant variance, that is to say:

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \tag{6-33}$$

Assuming that there is only one independent variable, the homoskedasticity assumption means that the variability around of the regression line is the same for any value of *x*. In other words, variability does not increase or decrease when *x* varies, as shown in figure 2.7, part a) of chapter 2. In figure 6.1, a scatter plot is shown corresponding to a model in which disturbances are homoskedastic.

If the homoskedasticity assumption is not satisfied, then there is heteroskedasticity, or disturbances are heteroskedastic. In figure 2.7, part b) a model with heteroskedastic disturbances was represented: the dispersion increases with increasing values of *x*. Figure 6.2 shows the scatter diagram corresponding to a model in which the dispersion grows when *x* grows.

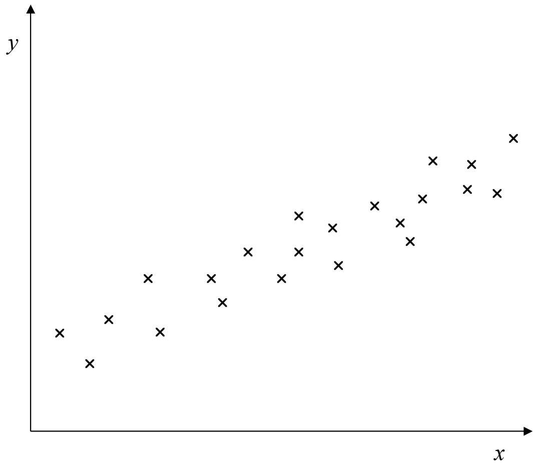


FIGURE 6.1. Scatter diagram corresponding to a model with homoskedastic disturbances.

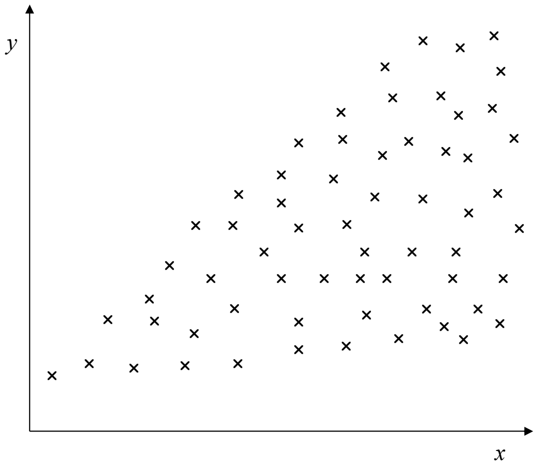


FIGURE 6.2. Scatter diagram corresponding to a model with heteroskedastic disturbances.

6.5.1 Causes of heteroskedasticity

In models estimated with cross sectional data (for example, demand studies based on surveys of household budgets) there are often problems of heteroskedasticity. However, heteroskedasticity can also occur in models estimated with time series.

Let us now consider some factors that can cause disturbances to be heteroskedastic:

a) Influence of the size of an explanatory variable in the size of the disturbance. Let us examine this factor using an example. Consider a model in which spending on hotels is a linear function of disposable income. If you have a representative sample of the population of a country, the great variability of the income received by families can be seen. Logically, low income families are unlikely to spend large amounts on hotels, and in this case we can expect that the oscillations in the expenditure of one family to another are not significant. In contrast, in high-income families a greater variability in this type of expenditure can be expected. Indeed, high-income families may choose between spending a substantial part of their income on hotels or spending virtually nothing. The scatter diagram in figure 6.2 may be adequate to represent what happens in a model to explain the demand for a luxury good such as spending on hotels.

b) The presence of outliers can cause heteroskedasticity. An outlier is an observation generated apparently by a different population to that generating the remaining sample observations. When the sample size is small, the inclusion or exclusion of such an observation can substantially alter the results of regression analysis and cause heteroskedasticity.

c) Data transformation. As we saw in a previous section, one of the solutions to solve the problem of multicollinearity consisted in transforming the model taking ratios with respect to a variable (say x_{ji}), i.e. dividing both sides of the model by x_{ji}. Therefore, the disturbance will now be u_i/x_{ji}, instead of u_i. Assuming that u_i fulfills the homoskedasticity assumption, the disturbances of the transformed model (u_i/x_{ji}) will no longer be homoskedastic but heteroskedastic.

6.5.2 Consequences of heteroskedasticity

When there is heteroskedasticity, the *OLS* method is not the most appropriate because the estimators obtained are not the *best*, i.e. the estimators are not *BLUE*.

Moreover, the *OLS* estimators obtained when there is heteroskedasticity, in addition to not being *BLUE*, have the following problem. The covariance matrix of the estimators obtained by applying the usual formula is not valid when there is heteroskedasticity (and/or autocorrelation). Consequently, the *t* and *F* statistics based on the estimated covariance matrix can lead to erroneous inferences.

6.5.3 Heteroskedasticity tests

We are going to examine two heteroskedasticity tests: Breusch-Pagan-Godfrey and White. Both of them are asymptotic and have the form of a Lagrange multiplier (*LM*) test.

Breusch-Pagan-Godfrey (BPG) test

Breusch and Pagan (1979) developed a test for heteroskedasticity and Godfrey (1978) developed another one. Because they are similar, they are usually known as Breusch–Pagan–Godfrey (*BPG*) heteroskedasticity tests.

The *BPG* test is an asymptotic test, that is to say, it is only valid for large samples. The null and alternative hypotheses of this test can be formulated as follows:

$$\begin{aligned} &H_0 : E\left(u_i^2\right) = \sigma^2 \quad \forall i \\ &H_1 : \sigma_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} \end{aligned} \tag{6-34}$$

where the z_i's can be some or all of the x_i's of the model.

Taking into account the above H_1, H_0 can be expressed as

$$H_0 : \alpha_2 = \alpha_3 = \cdots \alpha_m = 0 \tag{6-35}$$

The steps involved in this test are as follows:

Step 1. The original model is estimated and the *OLS* residuals are calculated.

Step 2. The following auxiliary regression is estimated, taking as the regressand the square of the residuals ($\hat{u}_i^2$) obtained in estimating the original model, since we know neither σ_i^2 nor u_i^2 :

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} + \varepsilon_i \qquad (6\text{-}36)$$

The auxiliary regression should have an intercept, although the original model is estimated without it. In accordance with expression (6-36), in the auxiliary regression there are *m* regressors in addition to the intercept.

Step 3. Designating by R_{ar}^2 the coefficient of determination of the auxiliary regression, the statistic nR_{ar}^2 is calculated.

Under the null hypothesis, this statistic (*BPG*) is distributed as follows:

$$BPG = nR_{ar}^2 \xrightarrow[n\to\infty]{} \chi_m^2 \qquad (6\text{-}37)$$

Step 4 For a significance level α, and designating by $\chi_m^{2(\alpha)}$ the corresponding value in χ^2 table, the decision to make is the following:

If $BPG > \chi_m^{2(\alpha)}$ H_0 is rejected

If $BPG \leq \chi_m^{2(\alpha)}$ H_0 is not rejected

In this test, high values of the statistic correspond to a situation of heteroskedasticity, that is to say, to the rejection of the null hypothesis.

EXAMPLE 6.5 Application of the Breusch-Pagan-Godfrey test

This test will be applied to a sub-sample of 10 observations, which have been used for estimating hotel expenditures (*hostel*) as a function of disposable income (*inc*). The data appear in table 6.5.

TABLE 6.5. *Hostel* and *inc* data.

i	*hostel*	*inc*
1	17	500
2	24	700
3	7	250
4	17	430
5	31	810
6	3	200
7	8	300
8	42	760
9	30	650
10	9	320

Step 1. Applying *OLS* to the model,

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

using data from table 6.5, the following estimated model is obtained:

$$\widehat{hostel}_i = \underset{(3.48)}{-7.427} + \underset{(0.0065)}{0.0533} inc_i$$

The residuals corresponding to this fitted model appear in table 6.6.

TABLE 6.6. Residuals of the regression of *hostel* on *inc*.

i	1	2	3	4	5	6	7	8	9	10
$\hat{u}_i$	-2.226	-5.888	1.100	1.505	-4.751	-0.234	-0.565	8.913	2.777	-0.631

Step 2. The auxiliary regression which must be estimated is the following:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 inc_i + \eta_i$$

Applying *OLS,* the following results are obtained:

$$\hat{u}_i^2 = -23.93 + 0.0799 inc \qquad R^2=0.5045$$

Step 3. Using the value of R^2, the *BPG* statistics is:

$$BPG= nR_{ar}^2 =10(0.56)=5.05.$$

Step 4. Given that $\chi_1^{2(0.01)}$=3.84, the null hypothesis of homoskedasticity is rejected for a significance level of 5%, because *BPG*>3.84, but not for the significance level of 1%.

Note that the validity of this test is asymptotic. However, the sample used in this example is very small.

White test

In the White test the hypothetical variables determining the heteroskedasticity are not specified. This test is a non-constructive test because it gives no indication of the heteroskedasticity scheme when the null hypothesis is rejected

The White test is based on the fact that the standard errors are asymptotically valid if we substitute the homoskedasticity assumption for the weaker assumption that the squared disturbance u^2 is uncorrelated with all the regressors, their squares, and their cross products. Taking this into account, White proposed to carry out the auxiliary regression of $\hat{u}_i^2$, since u_i^2 is unknown, on the factors mentioned above. If the coefficients of the auxiliary regression are jointly non-significant, then we can admit that the disturbances are homoskedastic. According to the assumption adopted, the White test is an asymptotic test.

The application of the White test can pose problems in models with many regressors. For example, if the original model has five independent variables, the White auxiliary regression would involve 16 regressors (unless some are redundant), which implies that the estimation is done with a loss of 16 degrees of freedom. For this reason, when the model has many regressors a *simplified* version of the White test is often applied. In the simplified version, the cross products are omitted from the auxiliary regression.

The steps involved in the *complete* version of the White test are as follows:

Step 1. The original model is estimated and the *OLS* residuals are calculated.

Step 2. The following auxiliary regression is estimated, taking as the regressand the square of the residuals obtained in the previous step:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2\psi_{2i} + \alpha_3\psi_{3i} + \cdots + \alpha_m\psi_{mi} + \varepsilon_i \tag{6-38}$$

In the above auxiliary regression, the regressors ψ_{ji} are the regressors of the original model, their squared values, and the crossproduct(s) of the regressors.

In any case, it is necessary to eliminate any redundancies that occur (i.e. regressors that appear repeatedly). For example, the intercept (which is 1 for all observations) and the square of the intercept cannot appear simultaneously as regressors, since they are identical. The simultaneous introduction of these two regressors will lead to perfect multicollinearity.

The auxiliary regression should have an intercept, even if the original model is estimated without it. In accordance with expression (6-38), in the auxiliary regression there are *m* regressors as well as the intercept.

Step 3. Designating by R_{ar}^2 the coefficient of determination of the auxiliary regression, the statistic nR_{ar}^2 is calculated.

Under the null hypothesis, this statistic (W) is distributed as follows:

$$W = nR_{ar}^2 \xrightarrow[n\to\infty]{} \chi_m^2 \tag{6-39}$$

This statistic is used to test the overall significance of model (6-38).

Step 4. It is similar to step 4 in Breusch-Pagan-Godfrey test.

EXAMPLE 6.6 *Application of the White test*

This test is going to be applied to data from table 6.5.

Step 1. This step is the same as in the Breusch-Pagan-Godfrey test.

Step 2. Since there are two regressors in the original model (the intercept and *inc*), the regressors of the auxiliary regression will be

$$\psi_{1i} = 1 \qquad \forall i$$
$$\psi_{2i} = 1 \times inc_i$$
$$\psi_{3i} = inc_i^2$$

Consequently, the model to be estimated is

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 inc_i + \alpha_3 inc_i^2 + \eta_i$$

By applying OLS to the data from table 6.5, we obtain the following

$$\hat{u}_i^2 = 14.29 - 0.10 inc_i + 0.00018 inc_i^2 \qquad R^2 = 0.56$$

Step 3. By using the $R^{2,}$ we obtain the W statistic:

$$W = nR^2 = 10(0.56) = 5.60.$$

The number of degrees of freedom is two.

Step 4. Given that $\chi_2^{2(0.10)}$=4.61, the null hypothesis of homoskedasticity is rejected for a 10% significance level because $W=nR^2$>4.61, but not for significance levels of 5% and 1%.

Note that the validity of this test is asymptotic too.

EXAMPLE 6.7 *Heteroskedasticity tests in models explaining the market value of the Spanish banks*

To explain the market value (*marktval*) of Spanish banks as a function of their book value (*bookval*) two models were formulated: one linear (example 2.8) and another one doubly logarithmic (example 2.10).

Heteroskedasticity in the linear model

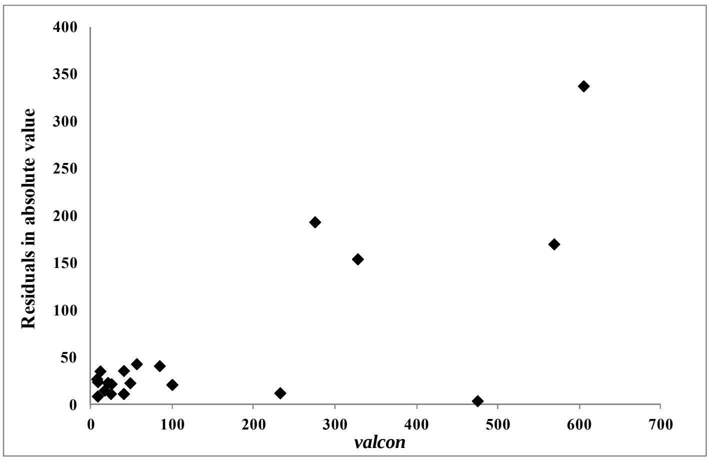
The linear model is given by

$$marktval = \beta_1 + \beta_2 bookval + u$$

Using data from 20 banks and insurance companies (filework *bolmad95*), the following results were obtained:

$$\widehat{marktval} = \underset{(30.85)}{29.42} + \underset{(0.127)}{1.219} bookval$$

In graphic 6.1, the scatter plot between the residuals in absolute value (ordinate) and the variable *bookval* (in abscissa) is represented. This graphic shows that the absolute values of the residuals, which are indicative of the spread of this series, grow with increasing values of the variable *bookval*. In other words, this graph provides an indication but not a formal proof of the existence of heteroskedasticity of the disturbances associated with the variable *bookval*.



GRAPHIC 6.1. Scatter plot between the residuals in absolute value and the variable *bookval* in the linear model.

The *BPG* statistic takes the following value:

$$BPG = nR_{ra}^2 = 20\times 0.5220 = 10.44$$

As $\chi_1^{2(0.01)}$ =6.64<10.44, the null hypothesis of homoskedasticity is rejected for a significance level of 1%, and therefore for α=0.05 and for α=0.10.

Now we will apply the White test. In this case, the auxiliary regression includes as regressors the intercept, the variable *bookval,* and the square of this variable. The White statistic takes the following value:

$$W = nR_{ra}^2 = 20\times 0.6017 = 12.03$$

As $\chi_2^{2(0.01)}$ =9.21=<12.03, the null hypothesis of homoskedasticity is rejected for a significance level of 1%.

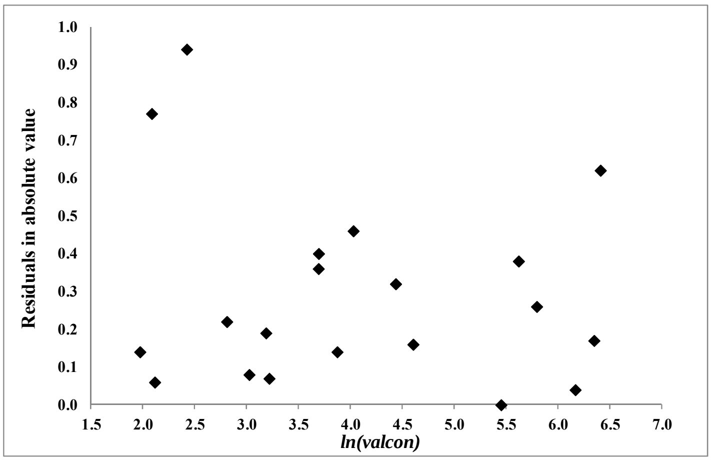
Therefore, both tests are conclusive in rejecting the null hypothesis for the usual levels of significance.

Heteroskedasticity in the log-log model

The estimated log-log model with the same sample was as follows:

$$\widehat{\ln(marktval)} = \underset{(0.265)}{0.676} + \underset{(0.062)}{0.9384}\ln(bookval)$$

In graphic 6.2 the scatter plot between the residuals in absolute value (ordinate), corresponding to this estimated model, and the variable ln(*bookval*) (in abscissa) is represented. As shown, the two largest residuals correspond to two banks with small market value. Even disregarding these two cases, apparently there is no relationship between the residuals and the explanatory variable of the model.



GRAPHIC 6.2. Scatter plot between the residuals in absolute value and the variable *bookval* in the log-log model.

The results of the two tests of heteroskedasticity applied are shown in table 6.7.

TABLE 6.7. Tests of heteroskedasticity on the log-log model to explain the market value of Spanish banks.

Test	*Statistic*	*Table values*
Breusch-Pagan	$BP = nR^2_{ra}$ =1.05	$\chi_2^{2(0.10)}$ =4.61
White	$W = nR^2_{ra}$ =2.64	$\chi_2^{2(0.10)}$ =4.61

Both tests carried out indicate that the null hypothesis of homoskedasticity cannot be rejected against the alternative hypothesis that the variance of the disturbances is associated with the explanatory variable of the model.

An important conclusion is that, if an econometric model is estimated with cross sectional data, it is easy to find observations with very different size. These problems of scale can cause heteroskedasticity in the disturbances but can often be solved by using log-log models.

EXAMPLE 6.8 Is there heteroskedasticity in demand of hostel services?

In general, heteroskedasticity in the disturbances does not usually appear in demand for food commodities. By contrast, heteroskedasticity is usually much more frequent in demand for luxury goods, because in the demand for these goods there is a large disparity in the behavior of high income households, while in households with low incomes such disparity is very unlikely.

In view of these considerations, the specification for analyzing the demand for hostel services is the following:

$$\ln\left(hostel\right) = \beta_1 + \beta_2 \ln(inc) + \beta_3 secstud + \beta_4 terstud + \beta_5 hhsize + u \qquad (6\text{-}40)$$

where *inc* is disposable income of a household, *hhsize* is the number of household members, and *secstud* and *terstud* are two dummies that take the value one if individuals have completed secondary and tertiary studies respectively.

The results obtained, using file *hostel*, are the following (file *hostel*):

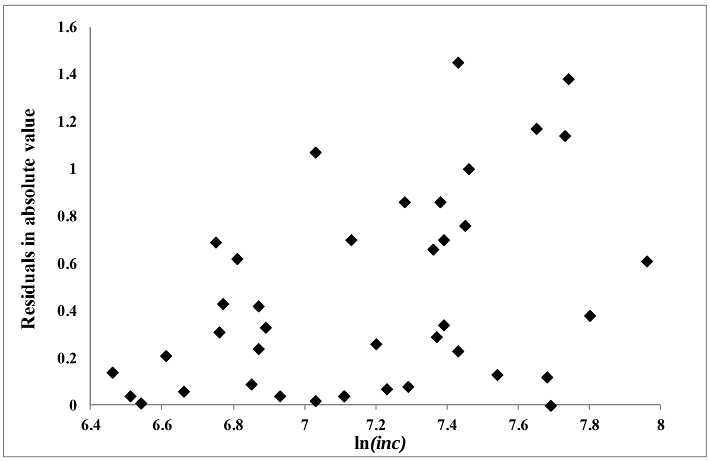
$$\widehat{\ln(hostel)}_i = \underset{(2.26)}{-16.37} + \underset{(0.324)}{2.732} \ln(inc)_i + \underset{(0.258)}{1.398} secstud_i + \underset{(0.333)}{2.972} terstud_i - \underset{(0.088)}{0.444} hhsize_i$$

$$R^2\text{=}0.921 \qquad n\text{=}40$$

Note that hostel services are a luxury good, as the elasticity of demand/income for this good is very high (2.73). This means that if income increases by 1%, spending on hostel services will increase, on average, by 2.73%. As can be seen, families where the main breadwinner has secondary studies (*secstud*) or, especially, higher education (*terstud*), spend more on hostel services than if the main breadwinner only has primary education. However, spending on hostel services will decrease as household size (*hhsize*) increases.

Graphic 6.3 shows the scatter plot between the residuals in absolute value and the variable ln(*inc*). Income (or a transformation of it) is the main candidate, if not the only one, to explain the hypothetical heteroskedasticity in the disturbances. As shown in the graphic, the dispersion of residuals is smaller for low incomes than for middle or upper incomes.

We will now apply the two tests of heteroskedasticity that have been discussed in this section.



GRAPHIC 6.3. Scatter plot between the residuals in absolute value and the variable *ln(inc)* in the hostel model.

The results of the two tests of heteroskedasticity applied are shown in table 6.8

TABLE 6.8. Tests of heteroskedasticity in the model of demand for hostel services.

Test	*Statistic*	*Table values*
Breusch-Pagan-Godfrey	$BPG = nR^2_{ra}$ =7.83	$\chi^{2(0.05)}_2$ =5.99
White	$W = nR^2_{ra}$ =12.24	$\chi^{2(0.01)}_2$ =9.21

In the *BPG* test we reject the null hypothesis of homoskedasticity for a significance level of α=0.05, but not for α=0.01.

Since there are many dummy variables in the model, including cross products in the auxiliary regression, this can lead to serious problems of multicollinearity. For this reason, in the auxiliary regression cross products are not included. Not surprisingly, among the regressors of the auxiliary regression squares of *secstud* and *terstud* are not included because they are dummies. Given the value obtained in the White statistic, we reject the null hypothesis of homoskedasticity for a significance level of α=0.01. Therefore, the White test is more conclusive in rejecting the homoskedasticity assumption.

6.5.4 Estimation of heteroskedasticity-consistent covariance

When there is heteroskedasticity and we apply *OLS*, we cannot make correct inferences by using the covariance matrix associated to the *OLS* estimates, because this matrix is not a consistent estimator of the covariance matrix of the coefficients. Consequently, the *t* and *F* statistics based on that estimated covariance matrix can lead to erroneous inferences.

Therefore, in the case that there is heteroskedasticity and *OLS* have been applied, a consistent estimate of the covariance matrix should be looked for to make inferences. White derived a consistent estimator of the covariance matrix under heteroskedasticity. However, it is important to note that this estimator does not work well if the sample is small, given that it is an asymptotic approximation.

Most econometric packages allow standard errors to be calculated by the White procedure. By using these consistent standard deviations, adequate tests can be made under the heteroskedasticity assumption.

EXAMPLE 6.9 Heteroskedasticity consistent standard errors in the models explaining the market value of Spanish banks (Continuation of example 6.7)

In the following estimated equation of the linear model, using file *bolmad95*, standard deviations of the estimates are calculated by the White procedure and therefore they are consistent under heteroskedasticity:

$$\widehat{marktval} = \underset{(18.67)}{29.42} + \underset{(0.249)}{1.219}\, bookval$$

As can be seen, the standard error of the *bookval* coefficient goes from 0.127 in the usual procedure to 0.249 in the White procedure. However, the *p*-value remains very low (0.0001). Accordingly, the significance of the variable *bookval* for all usual levels is still maintained. By contrast, the intercept, which has no special meaning in the model, now has a standard error (18.67), which is lower than that obtained with the usual procedure (30.85).

If we apply the White procedure to the log-log model, the following results are obtained:

$$\widehat{\ln(marktval)} = \underset{(0.3218)}{0.676} + \underset{(0.0698)}{0.9384}\ln(bookval)$$

In this case, the standard error of ln(*bookval*) coefficient is practically the same in the two procedures.

From the above results, the following conclusions can be obtained. In determining the market value of Spanish banks, disturbances of the linear model are strongly heteroskedastic. Therefore, when using a consistent estimate, the standard deviation is almost doubled compared to the standard one. By contrast, in the log-log model, which is not affected by heteroskedasticity, there is little difference between the standard errors obtained with both procedures.

6.5.5 The treatment of the heteroskedasticity

In order to estimate a model with heteroskedastic disturbances it is necessary to know or, if it is unknown, to estimate the pattern of heteroskedasticity. Thus, suppose that the standard deviation of the disturbances follows this scheme:

$$\sigma_i = f\left(x_{ji}\right) \tag{6-41}$$

As indicated in epigraph 6.1, the method *GLS* allows *BLUE* estimators to be obtained when disturbances are heteroskedastic. If we know scheme (6-41), the application of *GLS* is performed in two stages. In the first stage, the original model is transformed by dividing both sides by the standard deviation. Therefore, according to (6-41), the transformed model is given by

$$\frac{y_i}{f\left(x_{ji}\right)} = \beta_1 \frac{1}{f\left(x_{ji}\right)} + \beta_2 \frac{x_{1i}}{f\left(x_{ji}\right)} + \beta_3 \frac{x_{2i}}{f\left(x_{ji}\right)} + \cdots + \beta_k \frac{x_{ki}}{f\left(x_{ki}\right)} + \frac{u_i}{f\left(x_{ji}\right)} \tag{6-42}$$

It is easily seen that the disturbances of the previous model, ($u_i/f(x_{ji})$), are homoskedastic. Therefore, in the second stage *OLS* is applied to the transformed model, thus obtaining *BLUE* estimators. When we divide each observation by $f(x_{ji})$, we are weighting by the inverse of the value taken by this function. For this reason the above procedure is often called *weighted least squares* (*WLS*). In this case, the weighting factor is $1/f(x_{ji})$.

If the function $f(x_{ji})$ is not known, it is necessary to estimate it. In that case, the estimation method will not be exactly the *GLS* method because the application of this method involves the knowledge of the covariance matrix, or, at least, knowledge of a matrix that is proportional to it. If we estimate the covariance matrix, in addition to the parameters, it is said that *feasible GLS* is applied. In the case of heteroskedastic disturbances, the particularization of the feasible *GLS* method is called *WLS* (*weighted least squares*) in two stages. In the first the function $f(x_{ij})$ stage is estimated, whereas in the second stage *OLS* is applied to the model transformed using the $f(x_{ji})$ estimates.

To see how to apply the *WLS* method in two stages, let us consider the following relationship, which simply defines the variance of the disturbances, in the case of heteroskedasticity,

$$E\left(u_i^2\right)=\sigma_i^2 \tag{6-43}$$

Therefore, the squared disturbance can be made equal, as in the regression model, to its expectation plus a random variable. That is to say:

$$u_i^2=\sigma_i^2+\varepsilon_i \tag{6-44}$$

As the disturbances are not observable, one can establish a relationship analogous to the above using residuals instead of disturbances. Therefore,

$$\hat{u}_i^2=\sigma_i^2+\eta_{2i} \tag{6-45}$$

It should be noted that the above relationship does not have exactly the same properties as (6-44) because the residuals are correlated and heteroskedastic, even if the disturbances fulfill the *CLM* assumptions. However, in large samples they will have the same properties.

If we use the residuals as the regressand instead of the squared residuals, we must take the absolute values, since the standard deviation takes only positive values. Taking into account (6-45), the following relationship can be established:

$$|\hat{u}_i|=\sigma_i^2+\eta_{2i}=f\left(x_{ij}\right)+\eta_{2i} \tag{6-46}$$

Since the function *f(x*$_{ij}$*)* is generally unknown, different functions are often tried. Here there are some of the most common:

$$\begin{aligned}
|\hat{u}_i|&=\alpha_1+\alpha_2 x_{ji}+\eta_{2i}\\
|\hat{u}_i|&=\alpha_1+\alpha_2\sqrt{x_{ji}}+\eta_{2i}\\
|\hat{u}_i|&=\alpha_1+\alpha_2\frac{1}{x_{ji}}+\eta_{2i}\\
|\hat{u}_i|&=\alpha_1+\alpha_2\ln(x_{ji})+\eta_{2i}
\end{aligned} \tag{6-47}$$

The functional form with the best fit (a higher coefficient of determination or a smaller AIC statistic) is selected. For the transformation two circumstances are contemplated, depending on the significance of the intercept. If this coefficient is statistically significant, the model is transformed by dividing by the fitted values of the selected equation. If it is not statistically significant, the model is transformed by dividing by the regressor corresponding to the selected equation. Thus, if the selected equation were the second one of (6-47), with the intercept not being significant, the transformed model would be as follows:

$$\frac{y_i}{\sqrt{x_{ji}}}=\beta_1\frac{1}{\sqrt{x_{ji}}}+\beta_2\frac{x_{2i}}{\sqrt{x_{ji}}}+\beta_3\frac{x_{3i}}{\sqrt{x_{ji}}}+\cdots+\beta_k\frac{x_{ki}}{\sqrt{x_{ji}}}+\frac{u_i}{\sqrt{x_{ji}}} \tag{6-48}$$

Note that if the intercept is not significant, the estimated parameters are not involved in the transformation of the model, but they are if the intercept is significant. As the estimators in models (6-47) are biased, although consistent, it is not convenient to transform the models by applying the fitted values, $|\hat{u}_i|$-obtained by using $\hat{\alpha}_0$ and $\hat{\alpha}_1$- except when the significance of the intercept is very high (e.g., exceeding 1%).

***Example* 6.10 Application of weighted least squares in the demand of hotel services (Continuation of example 6.8)**

Since the two tests applied to the model to explain the cost of hotel services indicate that the disturbances are heteroskedastic, we apply the weighted least squares method to estimate the model (6-40).

First, we estimate the four models (6-47), using as the regressand the residuals $\left|\hat{u}_i\right|$ -in absolute value- obtained in the estimation of model (6-40) by *OLS*. The results are presented below:

$$\widehat{\left|\hat{u}_i\right|} = \underset{(0.143)}{0.0239} + \underset{(2.73)}{0.0003}inc \qquad R^2 = 0.1638$$

$$\widehat{\left|\hat{u}_i\right|} = \underset{(-1.34)}{-0.4198} + \underset{(2.82)}{0.0235}\sqrt{inc} \qquad R^2 = 0.1733$$

$$\widehat{\left|\hat{u}_i\right|} = \underset{(5.39)}{0.8857} - \underset{(-2.87)}{532.1}\frac{1}{inc} \qquad R^2 = 0.1780$$

$$\widehat{\left|\hat{u}_i\right|} = \underset{(-2.46)}{-2.7033} + \underset{(2.88)}{0.4389}\ln(inc) \qquad R^2 = 0.1788$$

In the above results, the *t*-statistic appears below each coefficient.

The functional form in which ln(*inc*) appears as a regressor is selected because it corresponds to the highest R^2 obtained. Since the coefficient of the independent term is not statistically significant at 1%, following the recommendation, *WLS* are applied taking 1/ln(*inc*) as the weighting variable. In estimating *WLS,* the following results were obtained:

$$\widehat{\ln(hostel)}_i = \underset{(2.15)}{-16.21} + \underset{(0.309)}{2.709}\ln(inc)_i + \underset{(0.247)}{1.401}secstud_i + \underset{(0.326)}{2.982}terstud_i - \underset{(0.085)}{0.445}hhsize_i$$

$$R^2\text{=}0.914 \qquad n\text{=}40$$

Compared to the *OLS* estimates of example 6.5, it can be seen that the differences are very small, which is indicative of the robustness of the model.

6.6 Autocorrelation

No autocorrelation, or *no serial correlation* assumption (assumption 8 of the *CLM*) states that disturbances with different subscripts are not correlated with each other:

$$E(u_iu_j) = 0 \qquad i \neq j \qquad \text{(6-49)}$$

That is, the disturbances corresponding to different periods of time, or to different individuals, are not correlated with each other. Figure 6.3 shows a plot corresponding to disturbances which are not autocorrelated. The *x* axis is time. As can be seen, disturbances are randomly distributed above and below the line 0 (theoretical mean of *u*). In the figure, each disturbance is linked by a line to the disturbance of the following period: in total this line crosses the line 0 on 13 occasions.

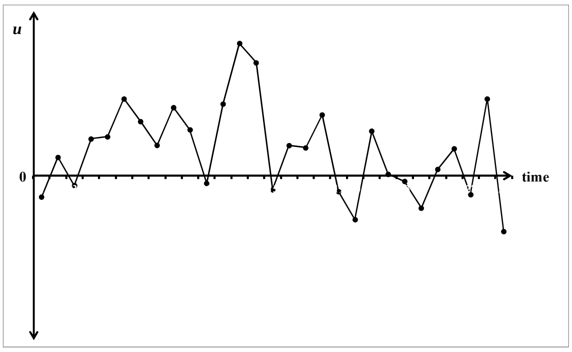


Figure 6.3. Plot of non-autocorrelated disturbances.

The transgression of the no autocorrelation assumption occurs quite frequently in models using time series data. It should be noted also that autocorrelation can be positive as well as negative. Positive autocorrelation is characterized by leaving a trail over time, because the value of each disturbance is near the value of the disturbance which precedes it. Positive autocorrelation occurs, by far, much more frequently in practice than the negative one. Figure 6.4 shows a plot corresponding to disturbances which are positively autocorrelated. As can be seen, the line which links successive disturbances crosses the line 0 only 4times.

By contrast, disturbances affected by negative autocorrelation present a saw tooth configuration, since each disturbance often takes the opposite sign of the disturbance which precedes it. In figure 6.5, the plot corresponds to disturbances which are negatively autocorrelated. Now the line 0 is crossed 21 times by the line which links successive disturbances.

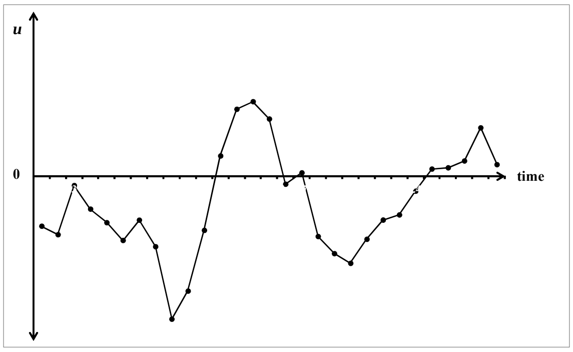


FIGURE 6.4. Plot of positive autocorrelated disturbances.

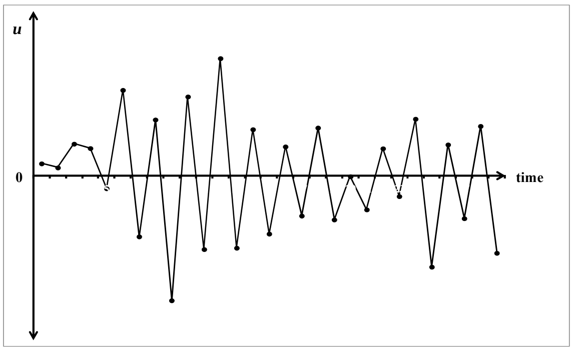


FIGURE 6.5. Plot of negative autocorrelated disturbances.

6.6.1 Causes of autocorrelation

There are several reasons for the presence of autocorrelation in a model, some of which are as follows:

a) Specification bias. That is, it can be caused by using an incorrect functional form or the omission of a relevant variable.

Let us suppose the correct functional form for determining *wage* as a function of years of experience (*exp*) is as follows:

$$wage = \beta_1 + \beta_2 exp + \beta_3 exp^2 + u$$

Instead of this model, the following one is fitted:

$$wage = \beta_1 + \beta_2 exp + v$$

In the second model, the disturbance has a systematic component ($v = \beta_3 exp^2 + u$). In figure 6.5, a scatter diagram (generated for the first model) and the fitted function of the second model are represented. As can be seen, for the low values of *exp* the fitted model overestimates wages; for intermediate values of *exp* wages are underestimated; finally, for high values the fitted model again overestimates wages. This example illustrates a case in which the use of an uncorrected functional form provokes positive autocorrelation.

On the other hand, the omission of a relevant variable in the model could induce positive autocorrelation if that variable has, for example, a cyclical behavior.

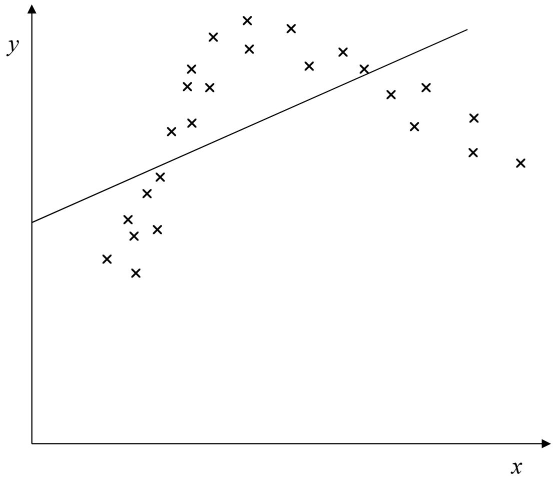


FIGURE 6.6. Autocorrelated disturbances due to a specification bias.

b) Inertia. The disturbance term in a regression equation reflects the influence of those variables affecting the dependent variable that have not been included in the regression equation. To be precise, inertia or the persisting effects of excluded variables of the model –and included in *u*- is probably the most frequent cause of positive autocorrelation. As is well known, macroeconomic time series -such as *GDP*, production, employment and price indexes- tend to move together: during expansion periods these series tend to increase in parallel, while in times of contraction they tend to decrease also in a parallel form. For this reason, in regressions involving time series data, successive observations of the disturbance are likely to be dependent on the previous ones. Thus, this cyclical behavior can produce autocorrelation in the disturbances.

c) Data Transformation. As an example let us consider the following model to explain consumption as a function of income:

$$cons_t = \beta_1 + \beta_2 inc_t + u_t \tag{6-50}$$

For the observation *t*-1, we can write

$$cons_{t-1} = \beta_1 + \beta_2 inc_{t-1} + u_{t-1} \tag{6-51}$$

If we subtract (6-51) from (6-50), we obtain

$$\Delta cons_t = \beta_2 \Delta inc_t + \Delta u_t \tag{6-52}$$

where $\Delta cons_t = cons_t - cons_{t-1}$, $\Delta inc_t = inc_t - inc_{t-1}$ and $v_t = \Delta u_t = u_t - u_{t-1}$.

The equation (6-50) is known as a *level form* equation, while the equation (6-52) is known as the *first difference form* equation. Both of them are used in empirical analysis. If disturbance in (6-50) is not autocorrelated, the disturbance in (6-52), which is equal to $v_t = u_t - u_{t-1}$, will be autocorrelated, because v_t and $v_{t\text{-}1}$ have a common element ($u_{t\text{-}1}$). In any case it should be noted the model (6-52), as specified, poses other econometric problems which will not be addressed here.

6.6.2 Consequences of autocorrelation

The consequences of autocorrelation for *OLS* are somewhat similar to those of heteroskedasticity. Thus, if the disturbances are autocorrelated, then the *OLS* estimator is not *BLUE* because one can find an alternative unbiased estimator with smaller variance. In addition to not being *BLUE*, the estimator obtained by *OLS* under the assumption of autocorrelation presents the problem that the estimation of the covariance matrix of the estimators calculated by the *OLS* usual formulas is biased. Consequently, the *t* and *F* statistics based on this covariance matrix can lead to erroneous inferences.

6.6.3 Autocorrelation tests

In order to test autocorrelation, a scheme of autocorrelation of disturbances in the alternative hypothesis must be defined. We will examine three of the best known tests. In two of them (the Durbin and Watson test and Durbin's *h* test) the alternative hypothesis is a first-order autoregressive scheme, while the third one, called the Breusch–Godfrey test, is a general test of autocorrelation applicable to higher-order autoregressive schemes.

Durbin and Watson test

The econometricians Durbin and Watson proposed the *d* test in 1950. *DW* is also used to refer to this statistic.

Durbin and Watson proposed the following scheme for the disturbances u_i:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \rightarrow NID(0, \sigma^2) \tag{6-53}$$

The proposed scheme for u_t is a first-order autoregressive scheme, since the disturbances appear as regressand and also as regressor lagged a period. In the terminology of time series analysis, the scheme (6-53) is called *AR*(1), that is to say, an autoregressive process of order 1. The coefficient of this scheme is ρ, which is required to be less than 1 in absolute value so that the disturbances do not have an explosive character, when *n* grows indefinitely. The variable ε_t is a random variable with a normal and independent distribution (which means *NID*) with mean 0 and variance σ^2. Consequently, the variable ε_t fulfills the same assumptions as u_t in the *CLM* assumptions. The variables with these properties are often called white noise variables.

According to the sign of ρ being positive or negative, the autocorrelation will be positive or negative. On the other hand, almost always one-tailed test is performed, namely the alternative hypothesis is taken as either positive autocorrelation or negative autocorrelation.

The problem of constructing an autocorrelation test is that the disturbances are not observable. The test must therefore be based on the residuals obtained from the *OLS* estimation. This raises problems, since, under the null hypothesis that disturbances are not autocorrelated, residuals are autocorrelated. In the construction of their test, Durbin and Watson took these factors into account.

Let us now apply this test. Taking as a reference the scheme defined in (6-53), Durbin and Watson formulate the following null and alternative hypothesis of positive autocorrelation

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned} \tag{6-54}$$

Thus, u_t=ε_t is verified under the null hypothesis, i.e. the model fulfills the *CLM* assumptions.

The statistic used by Durbin and Watson for testing hypotheses (6-54) is the *d* or *DW* statistic, defined as follows:

$$d = DW = \frac{\sum_{t=2}^{n}\left(\hat{u}_t - \hat{u}_{t-1}\right)}{\sum_{t=1}^{n}\hat{u}_t^2} \qquad (6\text{-}55)$$

The statistical distribution of *d*, which is symmetrical with a mean equal to 2, is very complicated, since it depends on the particular form of the matrix of regressor **X**, the sample size (*n*) and the number of regressors (*k*) excluding the intercept.

However, for different levels of significance, Durbin and Watson obtained two values (d_L and d_U) for each value of *n* and *k*. The rules to test positive autocorrelation are:

If $d < d_L$, there is positive autocorrelation.
If $d_L \leq d \leq d_U$, the test is not conclusive. (6-56)
If $d > d_U$, there is not positive autocorrelation.

As can be seen, there are values where the test is not conclusive. This is due to the effect that the particular configuration of the matrix **X** has on the distribution of *d*.

If you want to test negative autocorrelation, the alternative hypothesis is the following:

$$H_1 : \rho < 0 \qquad (6\text{-}57)$$

In order to apply the negative autocorrelation test, it is taken into account that the statistic *d* has a symmetrical distribution ranging between 0 and 4. The rules, therefore, are the following:

Si $d > 4 - d_L$, there is negative autocorrelation.
Si $4 - d_U \leq d \leq 4 - d_L$, the test is not conclusive. (6-58)
Si $d < 4 - d_U$, there is not positive autocorrelation.

The Durbin and Watson test is not applicable if there are lagged endogenous variables as regressors.

To be applied to quarterly data, Wallis considered a fourth-order autoregressive scheme:

$$u_t = \rho_4 u_{t-4} + \varepsilon_i \qquad \left|\rho_4\right| < 1 \qquad \varepsilon_t \rightarrow NID(0, \sigma^2) \qquad (6\text{-}59)$$

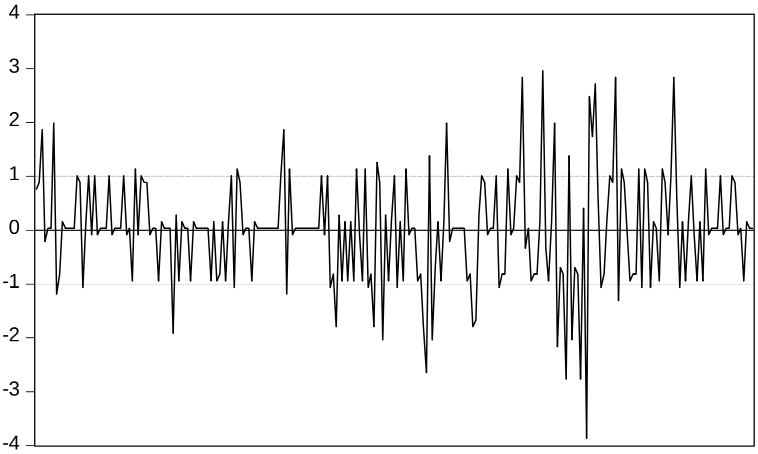
The above scheme is similar to (6-53), the difference being that the disturbance of the right hand side is lagged four periods. The Wallis statistic is similar to (6-55), but takes into account that the residuals are lagged four periods. This author designed *ad hoc* tables for testing models in which disturbances follow scheme (6-59).

EXAMPLE 6.11 Autocorrelation in the model to determine the efficiency of the Madrid Stock Exchange

In example 4.5, a model was formulated to determine the efficiency of the Madrid stock exchange. Graphic 6.4 shows the standardized residuals[2] corresponding to the estimation of this model, using file *bolmadef*. The *DW* statistic is equal to 2.04. (The *DW* statistic appears in the output of any econometric package). As the *DW* table does not have values for a sample size of 247, we use the corresponding values to n=200 and k'= 1. (In the nomenclature of this test, k' is used for the total number of regressors excluding the intercept). As the sample size is large we use a significance level of 1%. Upper and lower tabulated values, which correspond to the above specification, are as follows:

$$d_L=1.664; \qquad d_U=1.684$$

Since DW=2.04>d_U, we do not reject the null hypothesis that the disturbances are not autocorrelated for a significance level of α=0.01, i.e. of 1%, versus the alternative hypothesis of positive autocorrelation according to the scheme (6-53).



GRAPHIC 6.4. Standardized residuals in the estimation of the model to determine the efficiency of the Madrid Stock Exchange.

EXAMPLE 6.12 Autocorrelation in the model for the demand for fish

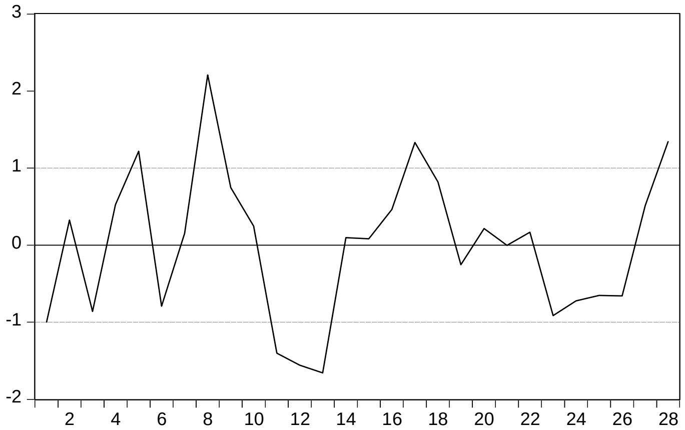
In example 4.9 we estimated model (4-44), using file *fishdem*, to explain the demand for fish in Spain. The graphic 6.5 shows the standardized residuals obtained in the estimation of this model. This graph does not show that there is a significant autocorrelation scheme. In this regard, it should be noted that, over a total of 28 observations, the line joining the points of the residuals crosses the axis 0 11 times, which indicates a degree of randomness of the distribution of the residuals.

The value of the *DW* statistic for testing the scheme (6-53) is 1.202. For n=28 and k'=3, and for a significance level of 1%, we get the following tabulated values:

$$d_L=0.969 \qquad d_U=1.415$$

Since d_L<1.202<d_U, there is not enough evidence to accept the null hypothesis, or to reject it.

[2] Standardized residuals are equal to residuals divided by $\hat{\sigma}$.



GRAPHIC 6.5. Standardized residuals in the model on the demand for fish.

Durbin's h test

Durbin (1970) proposed a statistic, called *h*, to test the hypothesis (6-54) in the case that one or more lagged endogenous variables appear as explanatory variables. The expression of the *h* statistic is the following:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n\widehat{\text{var}}\left(\hat{\beta}_j\right)}} \tag{6-60}$$

where $\hat{\rho}$ is the correlation coefficient between $\hat{u}_i$ and $\hat{u}_{i-1}$, n is the sample size, and $\widehat{\text{var}}\left(\hat{\beta}_j\right)$ is the variance corresponding to the coefficient of the lagged endogenous variable.

The statistic $\hat{\rho}$ can be estimated using the following approximation, $d \simeq 2(1-\hat{\rho})$. If the regressand appears with different time lags as regressors, the variance corresponding to the regressor with the lowest lag is selected.

Under assumptions (6-54), the *h* statistic has the following distribution:

$$h \xrightarrow[n\to\infty]{} N(0,1) \tag{6-61}$$

The critical region is therefore in the tails of the standard normal distribution: the tail on the right for positive autocorrelation and the tail on the left for negative autocorrelation.

The statistic (6-60) cannot be calculated if $n\widehat{\text{var}}\left(\hat{\beta}_j\right) \geq 1$. In this case, Durbin proposed an alternative procedure to estimate an auxiliary regression: the residuals are taken as the regressand, the regressors are the same as those of the original model and the residuals also lagged a period. This procedure is a particular case of the Breusch–Godfrey test, which we will see next.

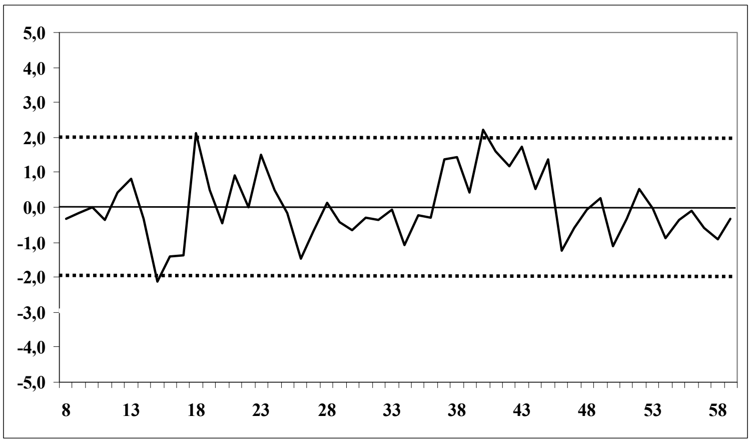
EXAMPLE *6.13 Autocorrelation in the case of Lydia E. Pinkham*

In example 5.5 with the case of Lydia E. Pinkham, a model to explain the sales of a herbal extract was estimated using file *pinkham*. Graphic 6.6 shows the graph of standardized residuals corresponding to this model. As can be seen, it appears that the residuals are not distributed in a random way. Note, for example, that from 1936 the residuals take positive values for 8 consecutive years.

The adequate test for autocorrelation in this model is Durbin's *h* statistic, as there is a lagged endogenous variable $sales_{t\text{-}1}$ in this model. The *h* statistic is:

$$h=\hat{\rho}\sqrt{\frac{n}{1-n\widehat{\operatorname{var}}\left(\hat{\beta}_j\right)}}=\left[1-\frac{d}{2}\right]\sqrt{\frac{n}{1-n\widehat{\operatorname{var}}\left(\hat{\beta}_j\right)}}=\left[1-\frac{1.2012}{2}\right]\sqrt{\frac{53}{1-53\times 0.0814^2}}=3.61$$

Given this value of h, the null hypothesis of no autocorrelation is rejected for α=0.01 or, even, for α=0.001, according to the table of the normal distribution.



GRAPHIC 6.6. Standardized residuals in the estimation of the model of the Lydia E. Pinkham case.

Breusch–Godfrey (BG) test

The Breusch–Godfrey (1978) test is a general test of autocorrelation applicable to higher-order autoregressive schemes, and it can be used when there are stochastic regressors such as the lagged regressand. This is an asymptotic test which is also known as the *LM* (Lagrange multipliers) general test for autocorrelation.

In the *BG* test, it is assumed that the disturbances u_t follow a pth-order autoregressive model $AR(p)$:

$$u_t=\rho_1 u_{t-1}+\rho_2 u_{t-2}+\cdots+\rho_p u_{t-p}+\varepsilon_t \qquad |\rho|<1 \qquad \varepsilon_t \rightarrow NID(0,\sigma^2) \qquad (6\text{-}62)$$

This is simply the extension of the $AR(1)$ scheme of the Durbin and Watson test.

The null hypothesis and the alternative hypotheses to be tested are:

$$H_0:\rho_1=\rho_2=\cdots=\rho_p=0$$
$$H_1:H_0 \text{ is not true}$$

The *BG* test involves the following steps:

Step 1. The original model is estimated and the *OLS* residuals ($\hat{u}_i$) are calculated.

Step 2. An auxiliary regression is estimated, in which the residuals ($\hat{u}_i$) are taken as the regressand and the regressors of the original model and the residuals lagged 1, 2, ... and p periods are taken as regressors:

$$\hat{u}_t=\alpha_1+\alpha_2 x_{2t}+\cdots+\alpha_k x_{kt}+\gamma_1\hat{u}_{t-1}+\cdots+\gamma_1\hat{u}_{t-p}+\varepsilon_i \qquad (6\text{-}63)$$

The auxiliary regression should have an intercept, even if the original model is estimated without it. In accordance with expression (6-63), in the auxiliary regression there are k+p regressors in addition to the intercept.

Step 3. Designating by R_{ar}^2 the coefficient of determination of the auxiliary regression, the statistic nR_{ar}^2 is calculated.

Under the null hypothesis, the *BG* statistic is distributed as follows:

$$BG = nR_{ar}^2 \xrightarrow[n\to\infty]{} \chi_{k+p}^2 \qquad (6\text{-}64)$$

The *BG* statistic is used to test the overall significance of the model (6-63). For this purpose, the *F* statistic can also be used. However, in this case it has only asymptotic validity, in the same way as with the *BG* statistic.

Step 4 For a significance level α, and designating by $\chi_{k+p}^{2(\alpha)}$ the corresponding value in χ^2 table, the decision to make is the following:

If $BG > \chi_{k+p}^{2(\alpha)}$ $\quad H_0$ is rejected

If $BG \leq \chi_{k+p}^{2(\alpha)}$ $\quad H_0$ is not rejected

As a particular case the *BG* test can be applied to quarterly data using a *AR*(4) scheme.

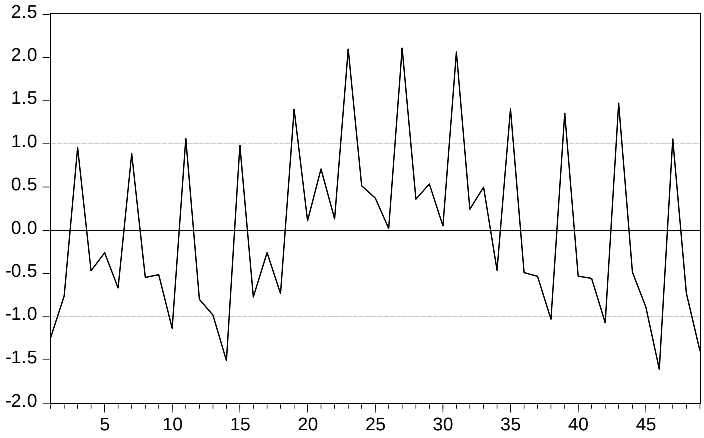
EXAMPLE 6.14 Autocorrelation in a model to explain the expenditures of residents abroad

To explain the expenditures of residents abroad (*turimp*), the following model was estimated by using quarterly data for the Spanish economy (file *qnatacsp*):

$$\widehat{\ln(turimp_t)} = \underset{(3.43)}{-17.31} + \underset{(0.276)}{2.0155}\ln(gdp_t)$$

R^2=0.531 $\qquad DW$=2.055 $\qquad n$=49

where *gdp* is gross domestic product.



GRAPHIC 6.7. Standardized residuals in the estimation of the model explaining the expenditures of residents abroad.

Graphic 6.7 shows the standardized residuals corresponding to this model. As can be seen, it appears that the residuals are not distributed in a random way because, for example, there are peaks every 4 quarters, indicating that the autocorrelation has a scheme *AR*(4).

The *BG* statistic, calculated for a *AR*(4) scheme, is equal to nR_{ar}^2 =36.35. Given this value of *BG*, the null hypothesis of no autocorrelation is rejected for α=0.01, since $\chi_5^{2(\alpha)}$=15.09. In the auxiliary regression, in which $\hat{u}_{t-1}, \hat{u}_{t-2}, \hat{u}_{t-3}$ and $\hat{u}_{t-4}$ have been used as regressors, $\hat{u}_{t-4}$ is the only significant regressor.

6.6.4 HAC standard errors

As an extension of White's heteroskedasticity-consistent standard errors that we have seen in section 6.5.2, Newey and West proposed a method known as *HAC* (heteroskedasticity and autocorrelation consistent) standard errors that allows *OLS* standard errors to be corrected not only in situations of autocorrelation, but also in the case of heteroskedasticity. Remember that the White method was designed specifically for heteroskedasticity. It is important to point out that the Newey and West procedure is, strictly speaking, valid in large samples and may not be appropriate in small ones. Note that a sample of 50 observations is a reasonably large sample.

EXAMPLE 6.15 HAC standard errors in the case of Lydia E. Pinkham (Continuation of example 6.13)

Given the existence of autocorrelation in the model for the case of Lydia E. Pinkham, we have calculated the standard errors according to the Newey and West procedure. These standard errors allow us to make hypothesis tests on parameters correctly. The available sample is 53 observations. In table 6.9 you can find the statistics *t* obtained by the conventional procedure and the procedure *HAC*, and the ratio between them. The *t* obtained by the procedure *HAC* are slightly lower than those obtained by the conventional method, except the *advexp* coefficient whose *t* is surprisingly much higher when the procedure HAC is applied. In any case, the same conclusions are obtained for the two methods for significance levels of 0.1, 0.05 and 0.01 in the significance test of each parameter.

TABLE 6.9.The *t* statistics, conventional and HAC, in the case of Lydia E. Pinkham.

regressor	*t* conventional	*t* HAC	ratio
intercept	2.644007	1.779151	1.49
advexp	3.928965	5.723763	0.69
sales(-1)	7.45915	6.9457	1.07
*d*1	-1.499025	-1.502571	1.00
*d*2	3.225871	2.274312	1.42
*d*3	-3.019932	-2.658912	1.14

6.6.5 Autocorrelation treatment

In order to estimate an econometric model where the disturbances follow the *AR*(1) scheme, we first consider the case that the value of ρ is known. Although this is more an academic assumption which would not happen in reality, it is convenient to adopt this assumption initially for presentation purposes. Let us suppose the following linear regression model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t \qquad (6\text{-}65)$$

If we lag a period in (6-65) and multiply both sides by ρ both, we obtain

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{2,t-1} + \rho\beta_3 x_{3,t-1} + \cdots + \rho\beta_k x_{k,t-1} + \rho u_{t-1} \qquad (6\text{-}66)$$

Subtracting (6-66) from (6-65), we have:

$$y_t - \rho y_{t-1} = \beta_1(1-\rho) + \beta_2\left(x_{2t} - \rho x_{2,t-1}\right) + \cdots + \beta_k\left(x_{kt} - \rho x_{k,t-1}\right) + \left(u_t - \rho u_{t-1}\right) \quad (6\text{-}67)$$

As can be seen, according to the scheme given in (6-53), the disturbance term of (6-67) fulfills the *CLM* assumptions.

Model (6-67) can be estimated directly by least squares if you know the value of ρ. The estimator obtained is close to the *GLS* method if the sample is large enough. The *GLS* method needs to strictly transform the observations 2 through *n* according to (6-67) scheme, but also to transform the first observation in the following way:

$$y_t\sqrt{1-\rho^2} = \beta_1\sqrt{1-\rho^2} + \beta_2\sqrt{1-\rho^2}\,x_{2t} + \cdots + \beta_k\sqrt{1-\rho^2}\,x_{kt} + \varepsilon_t \qquad (6\text{-}68)$$

When we estimate ρ together with the other model parameters, then the method is called *feasible GLS*.

In general, in the application of feasible *GLS* the transformation of the first observation according to (6-68) is ignored. Feasible *GLS* methods for estimating a model in which the disturbances follow a *AR*(1) scheme can be grouped into three blocks: a) two-step methods, b) iterative methods, and c) scanning methods.

Here we present two methods for block a), called direct method and Durbin two stages method.

In the first stage of these two methods, ρ is estimated. In the direct method, ρ is easily estimated from the *DW* statistic, using this approximate ratio $DW \simeq 2(1-\hat{\rho})$. In the method of Durbin in two stages, we estimate the following regression model in which the explanatory variables are the regressors of the original model, the regressors lagged one period and the endogenous variable lagged one period:

$$y_t = \alpha_1 + \alpha_{2,0}x_{2t} + \alpha_{2,1}x_{2,t-1} + \cdots + \alpha_{k0}x_{kt} + \alpha_{k1}x_{k,t-1} + \rho y_{t-1} + \upsilon_t \qquad (6\text{-}69)$$

The coefficient of the lagged endogenous variable is precisely the parameter ρ. In the first stage, the model (6-69) is estimated by *OLS*, taking from it the estimate of ρ. In the second stage, applicable to both methods, the model is transformed with the estimation of ρ calculated in the first stage as follows:

$$y_t - \hat{\rho}y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2\left(x_{2t} - \hat{\rho}x_{2,t-1}\right) + \cdots + \beta_k\left(x_{kt} - \hat{\rho}x_{k,t-1}\right) + \xi_t \qquad (6\text{-}70)$$

Applying *OLS* to the transformed model we obtain the parameter estimates. An exposition of iterative and scanning methods can be seen in Uriel, E.; Contreras, D.; Moltó, M. L. and Peiró, A. (1990): *Econometría. El modelo lineal.* Editorial AC. Madrid.

Exercises

Exercise 6.1 Let us consider that the population model is the following**:**

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad (1)$$

Instead, the following model is estimated:

$$\tilde{y}_i = \tilde{\beta}_2 x_{2i} \qquad (2)$$

Is $\tilde{\beta}_2$, obtained by applying *OLS* in (2), an unbiased estimator of β_3?

Exercise 6.2 Let us consider that the population model is the following:

$$y_i = \beta_2 x_i + u_i \qquad (1)$$

Instead, the following model is estimated:

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \qquad (2)$$

Is $\tilde{\beta}_2$, obtained by applying *OLS* in (2), an unbiased estimator of β_2?

Exercise 6.3 Let the following models be:

$$imp = \beta_1 + \beta_2 gdp + \beta_3 rpimp + u \qquad (1)$$

$$\ln(imp) = \beta_1 + \beta_2 \ln(gdp) + \beta_3 \ln(rpimp) + u \tag{2}$$

where *imp* is the import of goods, *gdp* is gross domestic product at market prices, and *rpimp* are the relative prices imports/gdp. The magnitudes *imp* and *gdp* are expressed in millions of pesetas.

a) Using a sample of the period 1971-1977 for Spain (file *imports*p), estimate models (1) and (2).

b) Interpret coefficients β_2 and β_3 in both models.

c) Apply the RESET procedure to model (1).

d) Apply the RESET procedure to model (2).

e) Choose the most adequate specification using the *p*-values obtained in sections *c)* and *d)*.

Exercise 6.4 Consider the following model of food demand

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

where *food* is spending on food, *rp* are the relative prices and *inc* is disposable income.

Researcher A omitted variable *inc*, obtaining the following estimation:

$$\widehat{food_i} = \underset{(11.85)}{89.97} + \underset{(0.118)}{0.107}\ rp_i$$

Researcher B, who is more careful, got the following estimation:

$$\widehat{food_i} = \underset{(5.84)}{92.05} - \underset{(0.067)}{0.142} rp_i + \underset{(0.031)}{0.236}\ inc_i$$

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

Throughout the discussion between researcher A and researcher B about which of the two estimated models is most appropriate, researcher A tries to justify his oversight on account of the omission being due to a problem of multicollinearity.

a) In favor of which researcher would you be in view of the results obtained? Explain your choice.

b) Obtain analytically the bias of the estimator of β_2 in the estimation carried out by researcher A.

Exercise 6.5 The following production function is formulated:

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u$$

where *output* is the amount of output produced, *labor* is the amount of labor, capital is the amount of capital.

The following data correspond to 9 companies:

$output_i$	230	140	180	270	300	240	230	350	120
$labor_i$	30	10	20	40	50	20	30	60	40
$capital_i$	160	50	100	200	240	190	160	300	150

A researcher estimates the model mistaking only 8 observations, and obtains the following results:

$$\widehat{output_i} = \underset{(1.956)}{97.259} + \underset{(0.124)}{0.970} labor_i + \underset{(0.027)}{0.650} capital_i$$

$$R^2 = 0.999 \qquad F=3422$$

The numbers in parentheses are the standard errors of the estimators and the *F* statistic corresponds to the test of the whole model.

When he realizes his mistake, he estimates the model with all observations (n=9), obtaining in this case the following results:

$$\widehat{output}_i = \underset{(32.046)}{75.479} - \underset{(1.742)}{1.970}\, labor_i + \underset{(0.376)}{1.272}\, capital_i$$

$$R^2 = 0.824 \qquad F = 14.056$$

His confusion is great when comparing the two estimates, and he cannot understand why the results become very different when using one more observation. Can we find any reason that could justify these differences?

Exercise 6.6 Suppose in the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

the R-squared obtained from regressing x_1 on x_2, which will be called $R^2_{1/2}$, is zero.

Run the following regressions:

$$y = \lambda_0 + \lambda_1 x_1 + u$$

$$y = \gamma_0 + \gamma_1 x_2 + u$$

a) Will $\hat{\lambda}_1$ be equal to $\hat{\beta}_1$ and $\hat{\gamma}_1$ be equal to $\hat{\beta}_2$?

b) Will $\hat{\beta}_0$ be equal to $\hat{\lambda}_0$ or $\hat{\beta}_0$ be equal to $\hat{\gamma}_0$?

c) Will var($\hat{\lambda}_1$) be equal to var($\hat{\beta}_1$) and var($\hat{\gamma}_1$) be equal to var($\hat{\beta}_2$)?

Exercise 6.7 An analyst wants to estimate the following model using the observations of the attached table:

$$y_i = e^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} x_{4i}^{\beta_4} e^{u_i}$$

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

What problems can occur in the estimation of this model with these data?

Exercise 6.8 In exercise 4.8, using the file *airqualy*, the following model was estimated:

$$\widehat{airqual}_i = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$

$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011}\, valadd_i$$

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

a) Calculate the statistic *VIF* for each coefficient.

b) What is your conclusion?

Exercise 6.9 To examine the effects of firm performance on CEO salary, the following model is formulated:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + \beta_6 age + u$$

where *roa* is the ratio profits/assets expressed as a percentage, *tenure* is the number of years as CEO (=0 if less than six months), and *age* is age in years. Salaries are expressed in thousands of dollars, and *sales* and *profits* in millions of dollars.

a) Using the full sample (447 observations) of the file *ceoforbes*, estimate the model by *OLS*.

b) Apply the normality test to the residuals.

c) Using the first 60 observations, estimate the model by *OLS*. Compare the coefficients and the R^2 of this estimation with that obtained in section *a)*. What is your conclusion?

d) Apply the normality test to the residuals obtained in section *c)*. What is your conclusion comparing this result with that obtained in section *b)*?

Exercise 6.10 Let the following model be

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

where

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

Apply generalized least squares to estimate β_2 in model [1].

Exercise 6.11 Let the following model be

$$y_i = \beta x_i + u_i \qquad [1]$$

where

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Estimate β in model [1] using generalized least squares.

b) Calculate the variance of the estimator of β.

Exercise 6.12 Let the model be

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

where the variance of the disturbances is equal to

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

1) Applying *OLS* to the model [1] and taking into account the Gauss-Markov assumptions, the variance of the estimator according to (2-16) is

$$\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad [2]$$

2) Applying *OLS* to the model [1] and considering that $\sigma_i^2 = \sigma^2 x_i$ and the remaining Gauss-Markov assumptions, the variance of the estimator is therefore equal to

$$\frac{\sigma^2 \sum (x_i - \bar{x})^2 x_i}{(\sum (x_i - \bar{x})^2)^2} \qquad [3]$$

3) Applying *GLS* to model [1] and considering that $\sigma_i^2 = \sigma^2 x_i$ and the remaining Gauss-Markov assumptions, the variance of the estimator is therefore equal to

$$\frac{\sigma^2}{\sum \frac{(x_i - \bar{x})^2}{x_i}} \qquad [4]$$

a) Are the variances [2] and [3] correct?

b) Show that [4] is less than or equal to [3]. (*Hint*: Apply the Cauchy-Schwarz inequality which says that $\left[\sum w_i z_i\right]^2 \leq \left[\sum w_i^2\right]\left[\sum z_i^2\right]$ is true)

Exercise 6.13 Let the following model be

$$hostel = \alpha_1 + \alpha_2 inc + u$$

where *hostel* is the spending on hotels and *inc* the yearly disposable income

The following information on 9 families was obtained:

family	*hostel*	*inc*
1	13	300
2	3	200
3	38	700
4	47	900
5	14	400
6	18	500
7	25	800
8	1	100
9	21	600

Hostel and income variables are expressed in thousands of pesetas.

a) Estimate the model by *OLS.*

b) Apply the White heteroskedasticity test.

c) Apply the Breusch-Pagan-Godfrey heteroskedasticity test.

d) Do you think it is appropriate to use the above heteroskedasticity tests in this case?

Exercise 6.14 With reference to the model seen in exercise 4.5, we assume now that

$$\text{var}(\varepsilon_i) = \sigma^2 \ln(y_i)$$

a) Are, in this case, the *OLS* estimators unbiased?

b) Are the *OLS* estimators efficient?

c) Could you suggest an estimator better than *OLS*?

Exercise 6.15 Indicate and explain which of the following statements are true when there is heteroskedasticity:

a) The *OLS* estimators are no longer *BLUE*.

b) The *OLS* estimators $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ are inconsistent.

c) The conventional *t* and *F* tests are not valid.

Exercise 6.16 In exercise 3.19, using the file *consumsp*, the Brown model was estimated for the Spanish economy in the period 1954-2010. The results obtained were the following:

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

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

Using the residuals of the above fitted model, the following regression was obtained:

$$\widehat{(\hat{u}_t^2)} = 141568 + 89.71 incpc_t - 149.2 conspc_{t-1}$$
$$-0.183 incpc_t^2 - 0.221 conspc_{t-1}^2 + 0.406 incpc_t \times conspc_{t-1}$$
$$R^2{=}0.285$$

a) Is there heteroskedasticity in the consumption function?

b) The following estimation, with White heteroskedasticity-consistent standard errors, is obtained:

$$\widehat{conspc_t} = \underset{(66.92)}{?} + \underset{(0.0669)}{?} incpc_t + \underset{(0.0741)}{?} conspc_{t-1}$$

Can you fill the blanks above? Please do so.

Explain the difference between the White heteroskedasticity- consistent standard errors and the usual standard errors of the initial equation.

c) Test whether the coefficient on *incpc* is equal to 0.5. What standard errors are you going to use in the inference process? Why?

Exercise 6.17 Assume the following specification:

$$c_i = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i$$
$$\sigma_i^2 = \sigma^2 h_i^2$$

Would it be appropriate to eliminate the heteroskedasticity to perform the following transformation?

$$\frac{c_i}{h_i} = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i \text{ ?}$$

Explain your answer.

Exercise 6.18 Let the following model be

$$y = \beta_1 + \beta_2 x + u$$

and we have the following information:

y_i	x_i	$\hat{u}_i$
2	-3	1.37
3	-2	-0.42
7	-1	0.79
6	0	-3.00
15	1	3.21
8	2	-6.58
22	3	4.63

a) Apply the White heteroskedasticity test.

b) Apply the Breusch-Pagan-Godfrey heteroskedasticity test.

c) Why is the significance obtained with both tests so different?

Exercise 6.19 Answer the following questions

a) Explain in detail what is the problem of heteroskedasticity in the linear regression model.

b) Illustrate briefly the problem of heteroskedasticity with an example.

c) Propose solutions to the heteroskedasticity problem.

Exercise 6.20 Using a sample corresponding to 17 regions, the following estimations were obtained:

$$\hat{y}_i = -309.8 + 0.76 z_i + 3.05 h_i \qquad R^2 = 0.989$$

$$\hat{u}_i^2 = -1737.2 - 17.8 z_i + 0.09 z_i^2 + 0.65 z_i h_i + 10.6 h_i - 0.31 h_i^2 \qquad R^2 = 0.705$$

where *y* is the expenditure on education, *z* is GDP and *h* is the number of inhabitants.

a) Is there a problem of heteroskedasticity? Detail the procedure followed in testing.

b) Assuming that the presence of heteroskedasticity is detected in the regression model, what solution would you take to test the significance of the explanatory variables of the model? Explain your answer.

Exercise 6.21 Using data from Spanish economy for the period 1971-1997 (file *importsp*), the following model was estimated to explain the Spanish imports (*imp*):

$$\widehat{\ln(imp_t)} = -\underset{(2.81)}{26.58} + \underset{(0.162)}{2.4336} \ln(gdp_t) - \underset{(0.021)}{0.4494} \ln(rpimp_t)$$

$$R^2=0.997 \quad n=27$$

where *gdp* is the gross domestic product at market prices, and *rpimp* are the relative prices imports/gdp. The variables *imp* and *gdp* are expressed in millions of pesetas.

a) Set up and estimate the auxiliary regression to perform the Breusch-Pagan-Godfrey heteroskedasticity test.

b) Apply the Breusch-Pagan-Godfrey heteroskedasticity test using the auxiliary regression run in section *a)*.

c) Set up the auxiliary regression to perform the *complete* White heteroskedasticity test.

d) Apply the *complete* White heteroskedasticity test using the auxiliary regression run in section *c)*.

e) Set up the auxiliary regression to perform the *simplified* White heteroskedasticity test.

f) Apply the *simplified* White heteroskedasticity test using the auxiliary regression run in section *e)*.

g) Compare the results of the test carried out in sections *b*), *d*) and *f*).

Exercise 6.22 Using data from file *tradocde*, the following model has been estimated to explain the imports (*impor*) in OECD countries:

$$\widehat{\ln(impor_i)} = \underset{(6.67)}{18.01} + \underset{(0.658)}{1.6425} \ln(gdp_i) - \underset{(0.636)}{0.5151} \ln(popul_i)$$

$$R^2=0.614 \qquad n=34$$

where *gdp* is gross domestic product at market prices, and *popul* is the population of each country.

a) What is the interpretation of the coefficient on ln(*gdp*)?

b) Set up the auxiliary regression to perform the White heteroskedasticity test.

c) Apply the White heteroskedasticity test using the auxiliary regression run in section *b)*.

d) Test whether the *import*/*gdp* elasticity is greater than 1. To make this test, do you need to use the White heteroskedasticity-robust standard errors?

Exercise 6.23 Explain in detail what the appropriate autocorrelation test would be in each situation:

a) When the model has no lagged endogenous variables and the observations are annual.

b) When the model has lagged endogenous variables and the observations are annual.

c) When the model has no lagged endogenous variables and the observations are quarterly.

Exercise 6.24 Two alternative models were used to estimate the average cost of annual car production of a particular brand in the period 1980-1999:

$$c=\alpha+\beta p+u \qquad R^2=0.848; \quad \bar{R}^2=0.812; \quad d=DW=0.51$$

$$c=\alpha+\beta p+\gamma p^2+u \qquad R^2=0.852; \quad \bar{R}^2=0.811; \quad d=DW=2.11$$

a) When comparing the two estimations, indicate if you detect any econometric problem. Explain it.

b) Depending on your answer to the previous section, which of the two models would you choose?

Exercise 6.25 In the period 1950-1980, the following production is estimated

$$\ln(o_t)=-\underset{(0.24)}{3.94}+\underset{(0.083)}{1.45}\ln(l_t)+\underset{(0.048)}{0.38}\ln(k_t)$$

$$R^2=0.994 \qquad DW=0.858 \qquad \hat{\rho}=0.559$$

where o is output, l is labor, and k is capital.

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

a) Test whether there is autocorrelation.

b) If the model had a lagged endogenous variable as an explanatory variable, indicate how you would test whether there is autocorrelation.

Exercise 6.26 Using 38 annual observations, the following demand function for a product was estimated:

$$d_i=2.47+\underset{(0.39)}{0.35}p_i+\underset{(0.06)}{0.9}d_{i-1} \qquad R^2=0.98 \qquad DW=1.82$$

where d is the quantity demanded, and p is the price.

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

a) Is there a problem of autocorrelation? Explain your answer.

b) List the conditions under which it would be appropriate to use the Durbin Watson statistic.

Exercise 6.27 The following model of housing demand with annual observations for the period 1960-1994 is estimated:

$$\widehat{\ln(rent_t)}=-\underset{(0.15)}{0.39}+\underset{(0.05)}{0.31}\ln(inc_t)-\underset{(0.02)}{0.67}\ln(price_t)+\underset{(0.04)}{0.70}\ln(rent_{t-1})$$

$$R^2=0.999 \qquad DW=0.52$$

where v is spending on rent, r is disposable income, p is the price of housing

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

a) Test whether there is autocorrelation.

b) Taking into account the conclusions reached in section *a),* how would you carry out the significance tests for each one of the coefficients? Explain your answer.

Exercise 6.28 Answer the following questions:

a) In a model to explain the sales, the estimation is carried out using quarterly data. Explain how you can reasonably test whether there is autocorrelation.

b) Describe in detail, introducing assumptions that you consider appropriate, how you would estimate the model when the null hypothesis of no autocorrelation is rejected.

Exercise 6.29 In the estimation of the Keynesian consumption function for the French economy, the following results were obtained:

$$\widehat{cons_t} = \underset{(0.73)}{1.22} + \underset{(79.39)}{0.854}\, inc_t$$

$$R^2 = 0.983 \quad \text{DW}=0.4205 \qquad n=30$$

(The numbers in parentheses are the *t* statistics of the estimators).

A researcher believes the focus should be placed on the saving function, rather than on the consumption function, proposing the following model:

$$saving_t = \alpha_1 + \alpha_2 inc_t + v_t$$

where

$$saving_t = inc_t - cons_t$$

a) Obtain the estimates of α_1 and α_2.

b) Estimate the variances of $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

c) Calculate the DW statistic of the saving model.

d) Calculate the R^2 of the saving model.

Exercise 6.30 Let the model be

$$\begin{gathered} y_t = \beta x_t + u_t \\ u_t = \rho u_{t-1} + \varepsilon_t; \quad E\left[\varepsilon_t^2\right] = \sigma^2 \qquad \forall i \end{gathered} \qquad [1]$$

a) If model [1] is transformed by taking differences first, under what circumstances is the transformed model preferable to model [1]?

b) Is it appropriate to use the R^2 to compare model [1] and the transformed model? Explain your answer.

Exercise 6.31 Let the model be:

$$y_t = \beta_1 + \beta_2 x_t + u_t \qquad [1]$$

The following sample of observations is disposable for the variables *x* and *y*:

y_i	6	3	1	1	1	4	6	16	25	36	49	64
x_i	-4	-3	-2	-1	1	2	3	4	5	6	7	8

a) Estimate the model [1] by OLS and calculate the corresponding adjusted determination coefficient.

b) Calculate the Durbin-Watson statistic for the estimations made in *a*).

c) In view of the Durbin and Watson test and the representation of the fitted line and residuals, is it appropriate to reformulate model [1]? Justify your answer and, if it is yes, estimate the alternative model that you consider the most appropriate for the data.

Exercise 6.32 Let the model be:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t; \qquad \varepsilon_t \sim NI\left(0, \sigma^2\right)$$

The following additional information is also disposable:

$$\rho = 0.5$$

y_i	22	26	32	31	40	46	46	50
x_i	4	6	10	12	13	16	20	22

a) Estimate the model by *OLS*.

b) Estimate the model by *GLS* without transforming the first observation.

c) Which of the two estimators of β_2 is more efficient?

Exercise 6.33 In a study on product demand, the following results were obtained:

$$\hat{y}_t = \underset{(7.17)}{2.30} + \underset{(0.05)}{0.86} x_t$$

$$R^2 = 0.9687 \quad \text{DW=3.4} \quad n = 15$$

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

Furthermore, the following additional information about the residual regressions is disposable:

1. $|\hat{u}_t| = \underset{(0.210)}{0.167} + \underset{(0.180)}{0.127} x_t$
2. $|\hat{u}_t| = \underset{(0.098)}{0.231} + \underset{(0.095)}{0.218} x_t^{1/2}$

a) Detect whether there is autocorrelation.

b) Detect whether there is heteroskedasticity.

c) What would be the most appropriate procedure to solve the potential problem of heteroskedasticity?

Exercise 6.34 Using a sample of the period 1971-1997 (file *importsp*), the following model was estimated, using *HAC* standard errors, to explain the imports of goods in Spain (*imp*):

$$\widehat{\ln(imp_t)} = -\underset{(3.65)}{26.58} + \underset{(0.210)}{2.434} \ln(gdp_t) - \underset{(0.023)}{0.4494} \ln(rpimp_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.73} \quad n = 27$$

where *gdp* is gross domestic product at market prices, and *rpimp* are the relative prices import/gdp. Both magnitudes are expressed in millions of pesetas.

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

a) Interpret the coefficient on *rpimp*.

b) Is there autocorrelation in this model?

c) Test whether the *imp/gdp* elasticity plus four times the *imp/rpimp* elasticity is equal to zero. (Additional information: $\text{var}(\hat{\beta}_2)$ =0.044247; $\text{var}(\hat{\beta}_3)$ =0.000540; and $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =0.004464).

d) Test the overall significance of this model.

Exercise 6.35 Using a sample for the period 1954-2009 (file *electsp*), the following model was estimated to explain the electricity consumption in Spain (*conselec*):

$$\widehat{\ln(conselec_t)} = \underset{(0.46)}{-9.98} + \underset{(0.035)}{1.469}\ln(gdp_t)$$

$$R^2 = 0.9805 \quad \text{DW=0.18} \quad n = 37 \tag{1}$$

where *gdp* is gross domestic product at 1986 market prices. The variable *conselec* is expressed in a thousand tonnes of oil equivalent (*ktoe*) and *gdp* is expressed in millions of pesetas.

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

a) Test whether there is autocorrelation applying the Durbin-Watson statistic.

b) Test whether there is autocorrelation applying the Breusch-Godfrey statistic for a *AR*(2) scheme.

c) The following model is also estimated:

$$\widehat{\log(conselec_t)} = \underset{(0.75)}{-0.917} + \underset{(0.107)}{0.164}\log(gdp_t) + \underset{(0.072)}{0.871}\log(conselec_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.93} \quad n = 36 \tag{2}$$

Test whether there is autocorrelation applying the procedure you consider appropriate.

d) Test whether the *conselec/gdp* elasticity in an equilibrium situation ($\ln(conselec^e) = \beta_1 + \beta_2 \ln(gdp^e) + \beta_3 \ln(conselec^e)$) is greater than 1, using an adequate procedure.

Exercise 6.36 The Phillips curve represents the relationship between the rate of inflation (*inf*) and the unemployment rate (*unemp*). While it has been observed that there is a stable short run tradeoff between unemployment and inflation, this has not been observed in the long run.

The following model reflects the Phillips curve:

$$inf = \beta_1 + \beta_2 unempl + u$$

Using a sample for the Spanish economy in the period 1970-2010 (file *phillipsp*), the following results were obtained:

$$\widehat{inf}_t = \underset{(1.79)}{12.59} - \underset{(0.120)}{0.3712}\,unempl_t$$

$$R^2\text{=0.198;} \qquad DW\text{=0.219;} \quad n\text{=41}$$

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

a) Interpret the coefficient on *unempl*.

b) Test whether there is first order autocorrelation using Durbin and Watson test.

c) Using the disposable information so far, can you test the significance of the coefficient on *unempl* adequately?

d) Using the *HAC* standard errors, test the significance of the coefficient on *unempl*.

Exercise 6.37 It is important to remark that the Phillips curve is a relative relationship. Inflation is considered low or high relative to the expected rate of inflation and unemployment is considered low or high relative to the so-called natural rate of unemployment. In the *augmented* Phillips curve this is taken into account:

$$inf_t - inf^e_{t/t-1} = \beta_2(unempl_t - \lambda_0) + u_t$$

where λ_0 is the natural rate of unemployment and $inf^e_{t/t-1}$ is the expected rate of inflation for t formed in t-1. If we consider that the expected inflation for t is equal to the inflation in t-1 ($inf^e_{t/t-1} = inf_{t-1}$) and $\beta_1 = -\beta_2\lambda_0$, the augmented Phillips curve can be written as:

$$inf_t - inf_{t-1} = \beta_1 + \beta_2 unempl_t + u_t$$

a) Using file *phillipsp*, estimate the above model.
b) Interpret the coefficient on *unempl*.
c) Test whether there is second order autocorrelation.
d) Test whether the natural rate of unemployment is greater than 10.

Appendix 6.1

First we are going to express the $\tilde{\beta}_2$ taking into account that y is generated by the model (6-8):

$$\begin{aligned}\tilde{\beta}_2 &= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)y_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(\beta_1+\beta_2 x_{1i}+\beta_3 x_{2i}+u_i)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{1i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \beta_3\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2 + \beta_3\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}\end{aligned} \tag{6-71}$$

If we take expectations on both sides of (6-71), we have

$$\begin{aligned}E(\tilde{\beta}_2) &= \beta_2 + \beta_3\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)E(u_i \mid x_2, x_3)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\ &= \beta_2 + \beta_3\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}\end{aligned} \tag{6-72}$$