

Thomas Andren

# Econometrics - Part III

Thomas Andren

---

# Econometrics

## Part III

---

BusinessSumup

---

Econometrics – Part III

© 2008 Thomas Andren & BusinessSumup

ISBN 978-87-7681-322-2

# Contents

<b>1.</b>	<b>Heteroskedasticity and diagnostics</b>	<b>6</b>
1.2	Detecting heteroskedasticity	7
1.2.1	Graphical methods	7
1.2.2	Statistical tests	10
1.3	Remedial measures	14
1.3.1	Heteroskedasticity-robust standard errors	19
<b>2.</b>	<b>Autocorrelation and diagnostics</b>	<b>22</b>
2.1	Definition and the nature of autocorrelation	22
2.2	Consequences	23
2.3	Detection of autocorrelation	26
2.3.1	The Durbin Watson test	27
2.3.2	The Durbins h test statistic	29
2.3.3	The LM-test	31
2.4	Remedial measures	32
2.4.1	GLS when AR(1)	32
2.4.2	GLS when AR(2)	33
<b>3.</b>	<b>Multicollinearity and diagnostics</b>	<b>35</b>
3.1	Consequences	35
3.2	Measuring the degree of multicollinearity	38
3.3	Remedial measures	41



**ANYTIME, ANYWHERE**

**LEARNING ABOUT  
SAP SOFTWARE HAS  
NEVER BEEN EASIER.**

SAP Learning Hub – the choice of  
when, where, and what to learn

**SAP** Learning Hub

**SAP**

<b>4.</b>	<b>Simultaneous equation models</b>	<b>43</b>
4.1	Introduction	43
4.2	The structural and reduced form equation	45
4.3	Identification	47
4.3.1	The order condition of identification	49
4.3.2	The rank condition of identification	50
4.4	Estimation methods	52
4.4.1	Indirect Least Squares (ILS)	52
4.4.2	Two Stage Least Squares (2SLS)	54
<b>A.</b>	<b>Statistical tables</b>	<b>58</b>
	Table A1	58
	Table A2	59
	Table A3	60
	Table A4	61



# 1. Heteroskedasticity and diagnostics

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

## 1.1 Consequences of using OLS

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

$$E[U_i] = 0 \quad (1.1)$$

$$V[U_i] = \sigma_U^2 \quad (1.2)$$

$$\text{Cov}[U_i, U_j] = 0 \quad (1.3)$$

Assumptions (1.1) and (1.3) are in use to make the OLS estimators unbiased and consistent. Assumption (1.2) is important for the OLS estimator to be efficient. Hence, if (1.2) is ignored we can no longer claim that our estimator is the best estimator among linear unbiased estimators. This means that it is possible to find another linear unbiased estimator that is more efficient.

### *Heteroskedasticity implies that*

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

It is important to understand that the violation of (1.2) makes the standard errors of the OLS estimators and the covariances among them biased and inconsistent. Therefore tests of hypothesis are no longer valid, since the standard errors are wrong. To see this, consider the variance of the estimator for the slope coefficient of the simple regression model:

$$V(b_1) = V \left[ \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 V[Y_i]}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \quad (1.4)$$

The expression given by (1.4) represents the correct variance that should be used. Unfortunately it involves the unknown population variance of the error term which is different for different observations.

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

$$V(b_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \neq \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = E[S_{b_1}^2] \quad (1.5)$$

As can be seen in (1.5) the variance of the OLS estimator is different from the expected value of the sample variance of the estimator that works under the assumption of a constant variance.

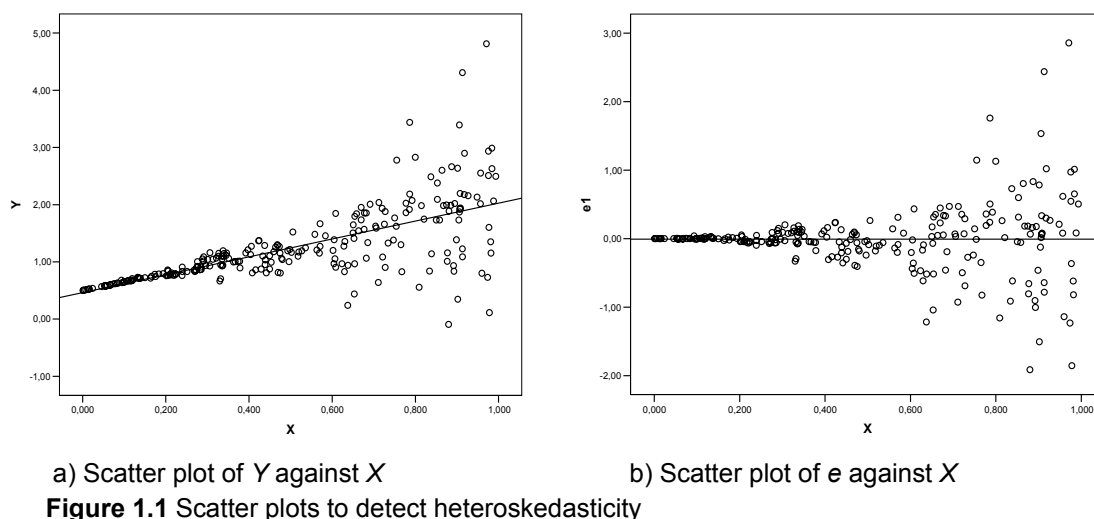
An important use of the regression equation is that of making predictions and forecasts of the future. Since the OLS estimators are unbiased and consistent, so will the forecasts. However, since the estimators are inefficient, the uncertainty of the forecasts will increase, and the confidence interval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

## 1.2 Detecting heteroskedasticity

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

### 1.2.1 Graphical methods

A natural starting point in detecting possible deviations from homoskedasticity is to plot the data. Since we are interested in the behavior of the error term and its variation, two obvious scatter plots are given in Figure 1.1, and that comes from a simple linear regression model. In Figure 1.1a the dependent variable is plotted against its explanatory variable  $X$ . Here we can see a clear pattern of heteroskedasticity which is driven by the explanatory variable. That is, the larger the value of  $X$ , the larger is the variance of the error term.



As an alternative to Figure 1.1a the estimated residuals could also be plotted directly against  $X$ , as is done in Figure 1.1b. In the simple regression case with just one explanatory variable these two graphs are always very similar. However, when using a multiple regression model the picture might be different, since the residual is a linear combination of all variables included in the model. Since it is the partial effect of  $X$  on the residual that is of primary interest, it is advised that the residual should be plotted against all involved variables separately. If it is possible to find a systematic pattern that gives indications of differences of the variances over the observations, one should be concerned. Graphical methods are useful, but sometimes it is difficult to say if heteroskedasticity is present and found harmful. It is therefore necessary to use statistical tests. The graphical method is therefore merely a first step in the analysis that can give a good picture of the nature of the heteroscedasticity we might have, which will be helpful later on when we are correcting for it.

### Example 1.1

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

$$\ln Y = B_0 + B_1 ED + B_2 ED^2 + B_3 year + B_4 year^2 + U \quad (1.6)$$

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

$$\ln \hat{Y} = 3.593 + 0.066ED - 0.001ED^2 + 0.017year - 0.000year^2 \quad (1.7)$$

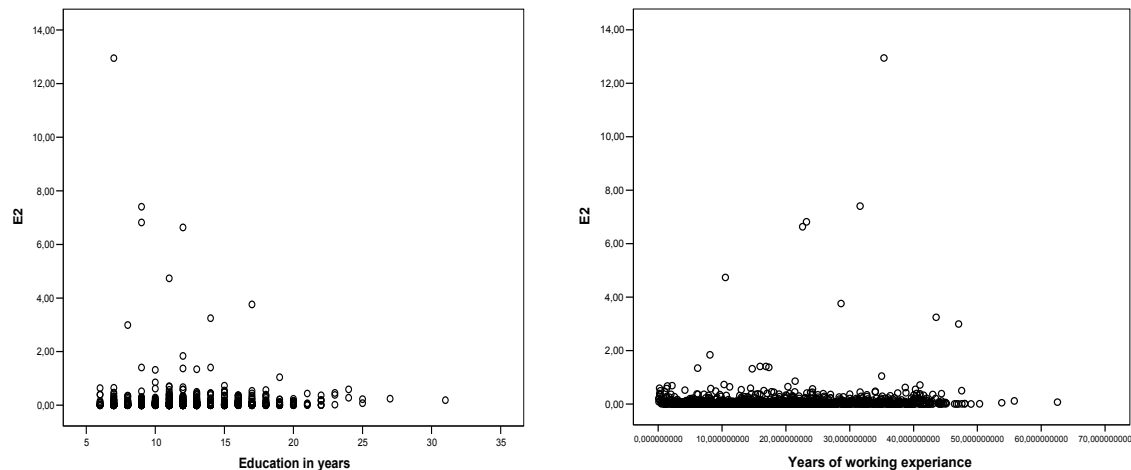
(40.1)    (5.2)            (−2.2)            (6.8)            (−2.9)

$\ln \hat{Y}$  should be interpreted as the predicted value of  $\ln Y$ . We observe that all coefficients are significantly different from zero. The squared terms have very small coefficients, even though their t-values are sufficiently large to make them significant. Observe that the coefficient for the square of  $year$  is different



from zero. Since the value is very small and we only report the first three decimals it appears to be zero. Its t-value shows that the standard error is even smaller.

We suspect that our residual might be heteroskedastic and we would like to investigate this by looking at a graph between the residual and the two explanatory variables. Sometimes, to enlarge the possible differences in variance among the residuals it is useful to square the estimated residual. If we do that we receive the graphs given in Figure 1.2.



a) Plot between  $e^2$  and  $ED$

b) Plot between  $e^2$  and  $year$

**Figure 1.2** Graphical analysis of error variance

# THE ANSWER TO YOUR LEARNING NEEDS

## GET QUALITY, FLEXIBLE, AND ECONOMICAL TRAINING WHEN AND WHERE IT'S NEEDED.

**SAP** Learning Hub



In Figure 1.2a we see the squared error term against the number of years of schooling, and a pattern can be identified. The error variance seems to be larger for lower years of schooling than for more years of schooling. This is of course just an indication that we need to investigate further using formal statistical test.

In Figure 1.1b the picture is less obvious. If ignoring the top outlier that makes it look like the variance is higher around 35 year of work experience, it is difficult to say if there is some heteroskedasticity to worry about. Since it is an unclear case, we still need to investigate the issue further, holding in mind that hypothesis testing is meaningless if the standard errors of the parameter estimates are wrong. Below we will go through the basic steps of the three most popular tests discussed in the textbook literature.

### 1.2.2 Statistical tests

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

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

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

$$RSS_1 = \sum_{i=1}^{n_1} e_i^2 \qquad RSS_2 = \sum_{i=n_1+1}^n e_i^2 \qquad (1.8)$$

- 3) Form the hypothesis that is to be tested:

$$\begin{aligned} H_0 : \sigma_i^2 &= \sigma^2 \\ H_1 : \sigma_i^2 &= \sigma^2 X_{1i} \end{aligned} \qquad (1.9)$$

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

$$F = \frac{S_1^2}{S_2^2} = \frac{RSS_1 / (n_1 - k)}{RSS_2 / (n_2 - k)} \sim F_{(n_1 - k, n_2 - k)} \qquad (1.10)$$

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

### Example 1.2

We are going to investigate model (1.6) used in Example 1.1 to see if we can identify any heteroskedasticity using the GQ-test. In the graphical analysis we found an indication of heteroskedasticity related to the number of years of schooling. Therefore, we sort the data set in an increasing order of years of schooling and delete the 33 percent in the middle. We use the two remaining sub-samples and estimate a regression for each sample. Using the results from these regressions we calculate the corresponding variance for each regression:

$$S_1^2 = \frac{RSS_1}{n_1 - k} = \frac{59.9641}{490} = 0.1223756 \quad S_2^2 = \frac{RSS_2}{n_2 - k} = \frac{43.6254}{489} = 0.08921339 \quad (1.11)$$

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

$$F = \frac{S_1^2}{S_2^2} = \frac{0.1223756}{0.0892133} = 1.3717 \quad (1.12)$$

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

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

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

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

$$E[U_i^2] = \sigma_i^2 = \sigma^2 f(A_0 + A_1X_1 + A_2X_2 + \dots + A_hX_h) \quad (1.13)$$

The variables included in (1.13) could be just a sub set of the explanatory variables of the model or it could be all of them. In Example 1.1 we could not be conclusive about whether just one or if both of our variables were driving the size of the variance. In a case like that it is advisable to included both the variables in the specification of the variance given by (1.13). The functional form is not expressed explicitly in (1.13) as stated, but we are going to use a linear specification, just as for the model we use.

The hypothesis of this test is:

$$\begin{aligned} H_0 : A_1 = A_2 = \dots = A_h = 0 \\ H_1 : A_j \neq 0 \text{ for at least one } j, j = 1, 2, \dots, h \end{aligned} \quad (1.14)$$

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

- 1) Run the regression for the model you believe suffers from heteroskedasticity using OLS.
- 2) Save the residuals and square them ( $e_i^2$ ). Use the squared residual and run the following auxiliary regression:

$$e^2 = A_0 + A_1X_1 + A_2X_2 + \dots + A_hX_h + \varepsilon \quad (1.15)$$

Equation (1.15) is a representation of (1.13) with a linear specification.



**MAXIMIZE PRODUCTIVITY**

**HELP YOUR ENTIRE ORGANIZATION BUILD EXPERTISE IN SAP SOFTWARE.**

**SAP Learning Hub**

**SAP**

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

$$LM = nR_e^2 \sim \chi_h^2 \quad (1.16)$$

where  $n$  is the number of observations used in the regression of (1.15) and  $R_e^2$  is the coefficient of determination received from (1.15). It turns out that the product of those two terms is chi-squared distributed with  $h$  degrees of freedom, where  $h$  is the number of restrictions, which in this case corresponds to the number of variables included in (1.15). The test value should therefore be compared with a critical value received from the Chi-square table for a suitable level of significance.

### Example 1.3

In this example we will use the same data set and the same model as in Example 1.2. But this time the test will involve both the variables included in the model. We choose not to include the squared terms, even though they in principle could be included. Following the basic procedure of the BP-test we specify and estimate the variance function with standard errors given within parenthesis:

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

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

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

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

**White's test** is very similar to the BP-test but does not assume any prior knowledge of the heteroskedasticity, but instead examines whether the error variance is affected by any of the regressors, their squares or cross products. Therefore, it is also a large sample test but it does not depend on any normality assumption. Hence, this third test is more robust than the other two test procedure described above, and is sometimes also called the White's General Heteroskedasticity test (WGH). The basic steps in the procedure are as follows for a model with two explanatory variables, where (1.17) represents the

main model and (1.18) the variance function that contains all the variables of the main function and their squares and cross products:

$$Y = B_0 + B_1X_1 + B_2X_2 + U \quad (1.17)$$

$$\sigma^2 = A_0 + A_1X_1 + A_2X_2 + A_3X_1^2 + A_4X_2^2 + A_5X_1X_2 + U \quad (1.18)$$

- 1) Estimate the parameters of equation (1.17) and create and save the residual.
- 2) Square the residual and run the auxiliary regression model given by (1.18).
- 3) Using the results from the auxiliary regression you can calculate the test value using (1.16). If the test value is larger than the critical value chosen, you reject the null hypothesis of homoskedasticity.

#### Example 1.4

We repeat the test executed in Example 1.3 and apply the WGH-test instead. Observe that the only difference is in the specification of the variance function. Following the basic steps given above we received the following results with standard errors reported within parenthesis:

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

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

Using these results we can calculate the test value:

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

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

### 1.3 Remedial measures

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

be expressed. Below we will look at some examples where we correct for heteroskedasticity under the assumption of a specific form of heteroskedasticity.

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

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

$$Y = B_0 + B_1X_1 + B_2X_2 + U \quad (1.19)$$

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

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

# FAST ADOPTION, FAST ROI

## EQUIP BUSINESS USERS TO ADOPT SAP SOLUTIONS.

SAP Learning Hub, user edition









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

$$\frac{Y_i}{\sqrt{X_{1i}}} = B_0 \frac{1}{\sqrt{X_{1i}}} + B_1 \frac{X_{1i}}{\sqrt{X_{1i}}} + B_2 \frac{X_{2i}}{\sqrt{X_{1i}}} + \frac{U_i}{\sqrt{X_{1i}}} \quad (1.20)$$

In practice this is carried out by just transforming  $Y$ ,  $X_1$ , and  $X_2$  and creating a new constant equal to  $1/\sqrt{X_{1i}}$ , instead of 1 that use to be there next to  $B_0$ . Hence when running this specification in a computer you have to ask the software to run the regression through the origin, since we now have a model specific constant that moves with  $X_1$ . All computer software has that option, and once you have found how to do that, you just regress  $\frac{Y_i}{\sqrt{X_{1i}}}$  on  $\frac{1}{\sqrt{X_{1i}}}$ ,  $\frac{X_{1i}}{\sqrt{X_{1i}}}$ ,  $\frac{X_{2i}}{\sqrt{X_{1i}}}$ ,  $\frac{U_i}{\sqrt{X_{1i}}}$ . When you transform the variables in this way, you automatically transform the error term, which now is divided by the square root of  $X_1$ . Once that is done, we have a homoskedastic error term. That is

$$V\left(\frac{U_i}{\sqrt{X_{1i}}}\right) = \left(\frac{1}{\sqrt{X_{1i}}}\right)^2 V(U_i) = \frac{1}{X_{1i}} \sigma^2 X_{1i} = \sigma^2 \quad (1.21)$$

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

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

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

$$V\left(\frac{U_i}{X_{1i}}\right) = \left(\frac{1}{X_{1i}}\right)^2 V(U_i) = \frac{1}{X_{1i}^2} \sigma^2 X_{1i}^2 = \sigma^2 \quad (1.22)$$

### Case 3: Two different variances

In this case we have an error term that takes only two values. Hence, our sample could include two groups with intrinsic differences in their variance. If these two groups are known, we can sort the data set with respect to these groups. For the first  $n_1$  observations, which contains the first group, the error term has the variance  $\sigma_1^2$  and for the remaining  $n_2$  observations, corresponding to the second group, the error term has the variance  $\sigma_2^2$ . In order to solve the heteroskedasticity problem here, we need to estimate the two



variances, by splitting the sample in two parts and estimate the regression variance separately for the two groups. Once that is done we proceed and transform as follows:

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

$$\begin{aligned} Y_i &= B_0 + B_1 X_{1i} + B_2 X_{2i} + U_i, \quad \text{var}(U_i) = \sigma_1^2 \quad i=1, \dots, n_1 \\ Y_i &= B_0 + B_1 X_{1i} + B_2 X_{2i} + U_i, \quad \text{var}(U_i) = \sigma_2^2 \quad i=n_1+1, \dots, n \end{aligned}$$

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

$$\begin{aligned} \frac{Y_i}{\sigma_1} &= B_0 \frac{1}{\sigma_1} + B_1 \frac{X_{1i}}{\sigma_1} + B_2 \frac{X_{2i}}{\sigma_1} + \frac{U_i}{\sigma_1} \quad i=1, \dots, n_1 \\ \frac{Y_i}{\sigma_2} &= B_0 \frac{1}{\sigma_2} + B_1 \frac{X_{1i}}{\sigma_2} + B_2 \frac{X_{2i}}{\sigma_2} + \frac{U_i}{\sigma_2} \quad i=n_1+1, \dots, n \end{aligned}$$

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

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

### Example 1.5

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

$$Y = B_0 + B_1 X_1 + B_2 X_2 + U$$

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

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

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

**Table 1.1** OLS and GLS estimates using 2000 observations

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

In Figure 1.3 we compare the residual plots before and after correcting for heteroskedasticity to see if the problem is fully solved. From Figure 1.3b the picture looks satisfying.

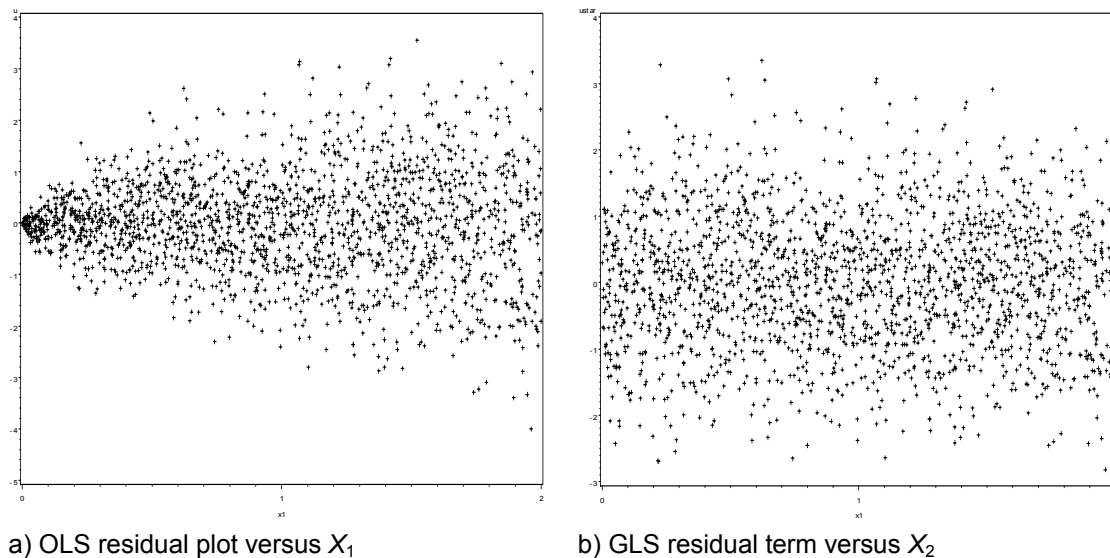
**Figure 1.3** Estimated residual plots before and after correction for heteroscedasticity

Table 1.1 show the results from OLS and GLS applied to a heteroskedastic model. As can be seen the estimated coefficient does not deviate that much which is what we expected since heteroskedasticity has no effect on the unbiasedness and consistency of the OLS estimator. When comparing the standard errors of the two estimations, large differences appear. The standard errors of the OLS estimated slope coefficients are twice as large as those for the corrected model. However the conclusions from the two estimations are the same. That is due to the relatively large sample that was used. If the sample would have been smaller, the corresponding t-values could have been much smaller and the chance of drawing the wrong conclusion would be greater. However, these results are sample specific. When the error term is heteroskedastic the standard errors are wrong and could be smaller or larger than the correct ones. So it is impossible to say something in advance without knowing something about the exact nature of the heteroskedasticity.

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

### 1.3.1 Heteroskedasticity-robust standard errors

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

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

$$V(b_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \quad (1.23)$$



**JUMP-START CAREERS**

**GIVE STUDENTS ONLINE  
ACCESS TO A VAST BODY  
OF KNOWLEDGE ABOUT  
SAP SOLUTIONS.**

SAP Learning Hub, student edition

**SAP** Learning Hub

**SAP**

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

$$S_{b_1}^2 \Big|_{Robust} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 e_i^2}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \quad (1.24)$$

Since (1.24) is a large sample estimator it is only valid asymptotically, and test based on them are not exact and when using small samples the precision of the estimator may be poor. Fortunately there exist a small sample adjustment factor that could improve the precision considerably by multiplying the variance estimator given by  $n/(n-k)$ . Furthermore and more importantly it is possible to generalize this formula to the multiple regression case, even though it become slightly more complicated. Fortunately most econometric software such as STATA and SAS, includes the option of receiving robust standard errors together with the parameter estimates when running the regression. Hence in the practical work of your own you should always use the robust standard errors when running regression models.

### Example 1.6

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

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

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

**Table 1.2** Regression results

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

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

Table 1.2 contains three regressions and the first column shows the results from the standard OLS regression assuming homoskedasticity. These results should be compared with the second column of estimates that use robust standard errors, which are heteroskedasticity consistent standard errors. Comparing those with the OLS case, we see that the robust standard errors are some what larger, which had consequences on the significance of the parameter for the squared education term, which no longer is significant. Including irrelevant variables in the regression makes the estimates less efficient. It therefore makes no sense to have the squared term included. In the third column, we re-estimate the model with out the squared term using robust standard errors.

Since we decided to use robust standard errors we could end up with a more parsimonious model, including only relevant terms. If we had included the squared education term, the marginal effects of education on earnings would be different and wrong. As can be seen from the RMSE measure that represents the estimated standard deviation of the error term it does not change very much among the specifications in Table 1.2. We should therefore conclude that the earnings model is not very sensitive to heteroskedasticity using this specification.



**LEARN BY DOING**

**DEVELOP EXPERTISE  
IN SAP SOLUTIONS  
THROUGH EXPLORATION  
AND PRACTICE.**

SAP Live Access

**SAP** Learning Hub



## 2. Autocorrelation and diagnostics

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

$$\text{Cov}[U_i, U_j] \neq 0 \quad \forall i \neq j \quad (2.1)$$

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

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

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

### 2.1 Definition and the nature of autocorrelation

An autocorrelated error term can take a range of different specifications to manifest a correlation between pair wise observations. The most basic form of autocorrelation is referred to as the first order autocorrelation and is specified in the following way:

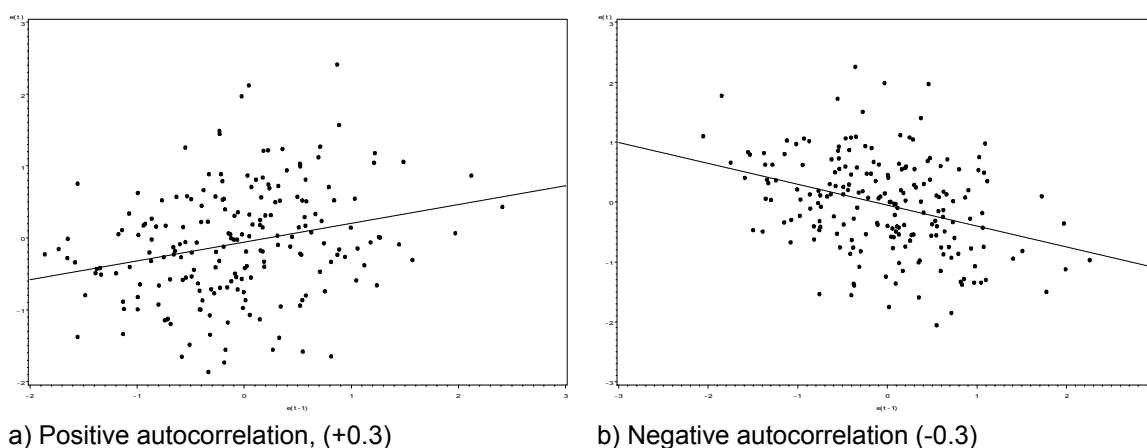
$$U_t = \rho U_{t-1} + V_t \quad (2.2)$$

where  $U$  refer to the error term of the population regression function. As can be seen from (2.2) the error term at period  $t$  is a function of it self in the previous time period  $t-1$  times the coefficient,  $\rho$ , which is referred to as the **first order autocorrelation coefficient** (This is the Greek letter rho, pronounced “row”). The last term  $V$ , is a so called white noise error term, and suppose to be completely random. It is often assume to be standard normal.

This type of autocorrelation is called autoregression because the error term is a function of its past values. Since  $U$  is a function of it self one period back only, as appose to several periods, we call it the first order autoregression error scheme, which is denoted AR(1). This specification can be generalized to capture up to  $n$  terms. We would then refer to it as the  $n$ th order of autocorrelation and it would be specified like this:

$$U_t = \rho U_{t-1} + \rho^2 U_{t-2} + \dots + \rho^n U_{t-n} + V_t \quad (2.3)$$

The first order autocorrelation is maybe the most common type of autocorrelations and is for that reason the main target of our discussion. The autocorrelation can be positive or negative, and is related to the sign of the autocorrelation coefficient in (2.2). One way to find out whether the model suffer from autocorrelation and whether it is positive or negative is to plot the residual term against its own lagged value.



**Figure 2.1** Scatter plots between  $e_t$  and  $e_{t-1}$

Figure 2.1 present two plots that are two examples of how the plots could look like when the error term is autocorrelated. The graph to the left represents the case of a positive autocorrelation with a coefficient equal to 0.3. A regression line is also fitted to the dots in order to make it easier to see in what direction the correlation drives. Sometimes you are exposed to plots where the dependent variable or the residual term is followed over time. However, when the correlation is below 0.5 in absolute terms, it might be difficult to identify any pattern using those plots, and therefore the plots above are preferable.

## 2.2 Consequences

The consequences of having an autocorrelated error term are very similar to those that appear with a heteroskedastic error term. In short we have that:



- 1) The estimated slope coefficients are unbiased and consistent.
- 2) With positive autocorrelation the standard errors are biased and too small.
- 3) With negative autocorrelation the standard errors are biased and too large.

Since the expected value of the residual is zero, despite any autocorrelation, the estimated slope coefficients are still unbiased. That is, the property of unbiasedness and consistency does not require uncorrelated error terms. Confirm this by reading chapter 3, book 1 where we derived the sample estimators and discussed their properties.

The efficiency property of the OLS estimator does, however, depend on the assumption of no autocorrelation. To see this, it is useful to repeat how the variance of the slope estimator looks like in the simple regression case. Assume the following set up:

$$Y_t = B_0 + B_1 X_t + U_t \quad (2.4)$$

$$U_t = \rho U_{t-1} + V_t \quad (2.5)$$

$$V_t \sim N(0,1) \quad (2.6)$$

$$E(U_t) = 0 \text{ and } V(U_t) = \sigma^2 \quad (2.7)$$

# HANDS-ON PRACTICE FOR EFFECTIVE LEARNING

## EXPERIENCE SAP SOFTWARE FIRSTHAND TO BUILD KNOWLEDGE AND ENHANCE SKILLS.

SAP Live Access

**SAP** Learning Hub





With this setup we observe that the residual term,  $U$ , is autoregressive of order one. The covariance is therefore given by:

$$\text{Cov}(U_t, U_{t-1}) = \text{Cov}(\rho U_{t-1}, U_{t-1}) = \rho \text{Cov}(U_{t-1}, U_{t-1}) = \rho \sigma^2 \quad (2.8)$$

When generalizing this expression to an arbitrary distance between two error terms it is possible to show that it equals

$$\text{Cov}(U_t, U_{t-j}) = \rho^j \sigma^2 \quad (2.9)$$

With this set up, together with the knowledge from chapter 3, book 1 on how the variance of the OLS estimator looks like, we can examine the variance under the assumption of autocorrelation. The variance of the slope coefficient can be expressed in the following way

$$\begin{aligned} V(b_1) &= \frac{1}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} V \left( \sum_{i=1}^n (X_i - \bar{X}) U_i \right) \\ &= \frac{\sigma^2}{\left( \sum_{t=1}^n (X_t - \bar{X})^2 \right)^2} \left( \sum_{t=1}^n (X_t - \bar{X})^2 + 2\rho \sum_{t=1}^n (X_t - \bar{X})(X_{t-1} - \bar{X}) + 2\rho^2 \sum_{t=1}^n (X_t - \bar{X})(X_{t-2} - \bar{X}) + \dots \right) \\ &= \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \left( 1 + 2\rho \frac{\sum_{t=1}^n (X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} + 2\rho^2 \frac{\sum_{t=1}^n (X_t - \bar{X})(X_{t-2} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} + \dots \right) \quad (2.10) \end{aligned}$$

If the autocorrelation coefficient were zero (*i.e.*  $\rho = 0$ ), the infinite series within the parenthesis in (2.10) would equal one. However, if ignoring the autocorrelation when present, we disregard this term which bias the variance of the slope coefficient. To receive a picture of how large in size the parenthesis is, it is useful to rewrite it into something more compact. In order to do that, we need to impose some assumptions on the behavior of  $X$ . We assume that the variance of  $X$  is constant and given by  $\sigma_X^2$ , and follows a first order autoregressive scheme, just like the error term of the model. This implies that  $\text{Cov}(X_t, X_{t-j}) = r^j \sigma_X^2$ , with  $r$  being the correlation coefficient for  $X_t$  and  $X_{t-1}$ . If we apply these assumptions to (2.10) we receive

$$V(b_1) = \frac{\sigma_U^2}{T \sigma_X^2} (1 + 2\rho r + 2\rho^2 r^2 + 2\rho^3 r^3 + \dots) = \frac{\sigma_U^2}{T \sigma_X^2} \frac{1 + \rho r}{1 - \rho r} \quad \text{when } j \rightarrow \infty \quad (2.11)$$

In order to receive the compact expression given by (2.11) you have to know how to deal with geometric series. If you do not know that, do not worry! The important thing here is to see how the sign of the autocorrelation coefficient and the correlation between of  $X_t$  and  $X_{t-j}$  affect the size of the variance and induce a bias when ignoring autocorrelation. We know that both  $\rho$  and  $r$  takes values between -1 and 1, since they represents correlation coefficients. With this set up we can analyze the size of the adjustment factor due to autocorrelation. Let us investigate two basic and common cases:

1)  $\rho > 0$  and  $r > 0$  (Positive autocorrelation)

When this is true, the adjustment factor become:  $\frac{1 + \rho r}{1 - \rho r} > 1$ , which means that the usual estimates for the

variance will be too small, and coefficients may appear more significant then they rely are. With a fixed value of  $r$ , the adjustment factor is increasing with the size of the autocorrelation coefficient, which increases the bias. If the value of  $r$  is zero, as it would be in cross sectional data for instance, the adjustment factor would be one, and the bias of the variance would be zero, independent of the size the autocorrelation coefficient. Most macro economic time series has an  $r$  value that is different from zero, and hence the case would in general not appear.

2)  $\rho < 0$  and  $r > 0$  (Negative autocorrelation)

With a negative autocorrelation, the adjustment factor become:  $0 < \frac{1 - \rho r}{1 + \rho r} < 1$ , which means that the usual

estimates will be too large, and appear less significant then they rely are. With a fixed value of  $r$ , and an increasing value of the autocorrelation coefficient in absolute terms, the adjustment factor will be smaller, and increase the bias.

Hence, when we have autocorrelation amongst our residual terms, we get biased estimates of the standard errors of the coefficients. Furthermore, the coefficient of determination and the usual estimator for the error variance of the model will be bias as well. Autocorrelation is therefore a serious problem that needs to be addressed.

## 2.3 Detection of autocorrelation

From the previous discussion we understand that autocorrelation is bad which emphasize the importance of learning how to detecting it. Below we will describe the most common procedures found in the text book literature. We will not discuss any graphical methods since they sometimes are difficult to interpret. In the introduction of the chapter we gave some examples on how graphical methods could be used. In more advanced time series analysis, graphical methods based on autocorrelation functions and partial autocorrelation functions are used frequently. However, we will not discuss these methods here.

### 2.3.1 The Durbin Watson test

The Durbin Watson test (DW) is maybe the most common test for autocorrelation and is based on the assumption that the structure is of first order. Since first order autocorrelation is most likely to appear in time series data, the test is very relevant, and all statistical software has the option of calculating it automatically for you.

The Durbin-Watson test statistic for first order autocorrelation is given by:

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T (e_t)^2} \quad (2.12)$$



**ANYTIME, ANYWHERE**

**LEARNING ABOUT  
SAP SOFTWARE HAS  
NEVER BEEN EASIER.**

SAP Learning Hub – the choice of  
when, where, and what to learn

**SAP** Learning Hub

**SAP**

with  $e$  being the estimated residual from a sample regression model. To see that this test statistic is related to the first order autocorrelation case we may rewrite (2.12) in the following way:

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T (e_t)^2} = \frac{\sum_{t=2}^T e_t^2}{\sum_{t=1}^T (e_t)^2} + \frac{\sum_{t=2}^T e_{t-1}^2}{\sum_{t=1}^T (e_t)^2} - \frac{2 \sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T (e_t)^2} \approx 1 + 1 + 2\hat{\rho} = 2(1 - \hat{\rho}) \quad (2.13)$$

where  $\hat{\rho}$  on the right hand side is the autocorrelation coefficient from an first order autoregression scheme. However, it is only an approximation since the expressions in the numerator sum from 2 to  $T$  instead of 1 to  $T$  as is the case in the denominator. The larger the value of  $T$  the better is the approximation.

From (2.13) it is possible to see that the  $DW$  test statistic only takes values between 0 and 4 since the autocorrelation coefficient only takes values between -1 and 1. Hence when the autocorrelation coefficient equals 0, the  $DW$  test statistics equals 2. If  $DW > 2$  we have an indication of a negative autocorrelation, and if  $DW < 2$  we would have an indication of a positive autocorrelation. However, since the relationship is an approximation, the  $DW$  test value can sometimes deviate from 2 even though the autocorrelation coefficient is zero. So the standard question is how much it is allowed to deviate? Could we use some critical values to help us interpret the estimated value of  $DW$ .

Unfortunately there exist no simple distribution function for this test function since it depends on the number of observations used as well as the values of the explanatory variables used in the regression. For that reason it is not possible to establish a precise critical value for the  $DW$  test statistic. However, Durbin and Watson made some simulations so that you, based on the number of observations used, and the number of parameters included in the model, can find a lower value (L) and an upper value (U) to compare the  $DW$  test value with.

**Table 2.1** Possible outcomes from the Durbin-Watson test for autocorrelation

Positive AC	Inconclusive	No AC	Inconclusive	Negative AC
$0 < DW < L$	$L < DW < U$	$U < DW < (4-U)$	$(4-U) < DW < (4-L)$	$(4-L) < DW < 4$

Table 2.1 show five different regions where the  $DW$ -test value potentially could end up. If you receive a test value that is located in the interval between the lower value (L) and the upper value (U) your test is inconclusive and you have no use of the  $DW$ -test. However, if the  $DW$ -value is between 0 and the lower value (L) you can draw the conclusion of having a positive autocorrelation. In the statistical table, with upper and lower values for the  $DW$ -test, you will only find the values that refer to the section below 2. In case of a negative autocorrelation you have to form the upper and lower value your self, using L and U as is done in Table 2.1.

**Example 2.1**

Assume that you have a time series with 150 observations, and two explanatory variables that will be used to explain the dependent variable. Running the regression you received a *DW*-test value equal to 1.63. In the table with critical values for the Durbin Watson test you found that  $L=1.71$  and  $U=1.76$ . Since the test value is outside the inconclusive interval and below the lower value we have to draw the conclusion that our model suffer from positive autocorrelation.

**Example 2.2**

We have the same set up as in the previous example but with a *DW*-test value equal to 2.35. In the table we found the value for  $L=1.71$  and the value for  $U=1.76$ . Using these values we can calculate the inconclusive interval related to *DW*-values larger than 2. Using the information in Table 2.1 we received:  $(4-U) = (4-1.76) = 2.24$  and  $(4-L) = (4-1.71) = 2.29$ . Since the test value of 2.35 is outside the interval and larger than the upper value of 2.29 we must conclude that our model suffer from negative autocorrelation.

**2.3.2 The Durbins h test statistic**

As been described above, the *DW*-test is made for the purpose of testing for first order autocorrelation. Furthermore, it assumes that none of the explanatory variables are lagged dependent variables which would be the case when estimating a dynamic model. When that is the case the *DW*-test has a tendency to be close to 2 even though the error terms are serially correlated. Hence, the *DW*-test should not be used with the following kind of model:

$$Y_t = B_0 + B_1 Y_{t-1} + B_2 X_t + U_t \quad (2.14)$$

Fortunately there is an easy alternative to the *DW*-test that could be seen as a modified version of it and for that reason is called the Durbins *h* statistic. It is defined in the following way:

$$h = \left(1 - \frac{DW}{2}\right) \sqrt{\frac{T}{1 - T[Var(b_1)]}} \quad (2.15)$$

where *DW* is the standard *DW*-test, *T* the number of observations and  $Var(b_1)$  the square of the standard error of the estimated parameter for the lagged dependent variable. The test statistic has been shown to be standard normally distributed under the null hypothesis of no autocorrelation, which means that the test value should be compared with a critical value from the standard normal table.

The presence of autocorrelation in models that include lagged dependent variables is even more affected than the standard model. When the error term is serially correlated in a dynamic model the estimated parameters are biased and inconsistent. It is therefore very important to correct for the problem before using the estimates for anything.

**Example 2.3**

Assume that we have estimated the parameters of a dynamic model and received the following results with standard errors within parenthesis:

$$\hat{Y}_t = 1.324 + 0.789Y_{t-1} + 0.821X_t \quad DW = 1.529$$

$$(0.321)(0.043) \quad (0.032)$$

We use quarterly data over a 30 years period, which means that  $T=120$ . Since our model includes a lagged variable we lose one observation in the estimation. Using the information from the regression, we may form the Durbins  $h$  statistic:

$$h = \left(1 - \frac{1.529}{2}\right) \sqrt{\frac{119}{1 - 119 \times (0.043)^2}} = 2.925$$

Using a one sided test at the 5 percent significance level we receive a critical value of 1.645. Since the test value is much larger than the critical value, we must conclude that our error terms are serially correlated.



### 2.3.3 The LM-test

The LM test is a more general test of autocorrelation compared to the two previous tests. Furthermore, it allows for a test of autocorrelation of higher order than one, and can be used even though lagged dependent variables are included in the model. However, the LM test is a large sample test which means that it should be treated as an approximation when using small samples, compared to the DW-test that could be seen as an exact test.

The LM test is executed with the following steps:

- 1) Estimate the parameters of your main model:  $Y_t = B_0 + B_1X_t + U_t$  (2.16)
- 2) Create the residual term using the estimated parameters and lag it.
- 3) Extend your original model by including the estimated lagged residual in the specification:  

$$Y_t = B_0 + B_1X_t + \rho e_{t-1} + V_t$$
 (2.17)
- 4) Test the null hypothesis  $H_0 : \rho = 0$  using a simple t-test. If you reject the null hypothesis you can conclude that you have autocorrelation.

The equation given by (2.17) can be extended to include more lags of the residual terms in order to test for higher order of autocorrelation.

#### Example 2.4

Assume that we have a time series model with two explanatory variables, and we suspect that the error term might be serially correlated of the second order.

$$Y_t = B_0 + B_1X_{1t} + B_2X_{2t} + U_t \quad (2.18)$$

Since the error term is unobserved we need to estimate the residuals for the model using the estimated parameters of the model:

$$e_t = Y_t - b_0 - b_1X_{1t} - b_2X_{2t} \quad (2.19)$$

Lag the estimated residuals from (2.19), re-specify equation (2.18) and receive:

$$Y_t = B_0 + B_1X_{1t} + B_2X_{2t} + \rho_1e_{t-1} + \rho_2e_{t-2} + V_t \quad (2.20)$$

We estimated the parameters of the extended version of the model given by (2.20) and received the following estimates with standard errors within parenthesis:

$$\hat{Y}_t = 3.241 + 0.378X_{1t} + 0.562X_{2t} + 0.324e_{t-1} + 0.124e_{t-2} \quad (2.21)$$

(1.301)(0.091)      (0.192)      (0.029)      (0.101)

By investigating the significance of the coefficients of the two residual terms, we see that the first one is significantly different from zero, while the other is not. We therefore conclude that the residual term is serially correlated of the first order only.

Since we included lagged variables in the specification we lose some observations, and when including two estimated residuals as in (2.21) we lose two. If we have a large sample, losing two observations is not a big deal. However, if we only have 20 observations, losing the first two observations might have an effect. One way to deal with this problem is to impose some values for  $e_0$  and  $e_{-1}$ . Since the expected value of the residual terms equal zero, the missing observations could be replaced by zeros. Running the regression with and without the imposed values will give you an indication if the two missing observations are important. Other more advanced methods might be available.

## 2.4 Remedial measures

Once we have found that our error term is serially correlated we need to correct for it before we can make any statistical inference on the population. As for the case of heteroskedasticity, we need to transform the involved variables and therefore use generalized least square. The transformation looks different dependent on the order of autocorrelation. We will therefore look at the two most frequently used error structures, AR(1) and AR(2), and show how it should be done for those two cases. After that it should be an easy task to generalize the transformation method for an autoregression of order  $n$ .

### 2.4.1 GLS when AR(1)

The transformation will be explained by an example. Assume that the objective is to transform the following model:

$$Y_t = B_0 + B_1 X_t + U_t \quad (2.22)$$

For simplicity reasons we use the specification of the simple regression model. However, the method can be generalized to any number of explanatory variables. The objective is to transform the autocorrelated  $U_t$  with something that free from autocorrelation  $V_t$ . Assume that  $U_t$  is autoregressive of order one and given by:

$$U_t = \rho U_{t-1} + V_t \quad (2.23)$$

If we substitute (2.23) into (2.22) we receive:

$$Y_t = B_0 + B_1 X_t + \rho U_{t-1} + V_t \quad (2.24)$$



Form the following expression using (2.22):

$$\rho U_{t-1} = \rho Y_{t-1} - \rho B_0 - \rho B_1 X_{t-1} \quad (2.25)$$

Substitute (2.25) into (2.24) and rearrange:

$$(Y_t - \rho Y_{t-1}) = B_0(1 - \rho) + B_1(X_t - \rho X_{t-1}) + V_t \quad (2.26)$$

Equation (2.26) is the transformed equation we are looking for. The error term of the original model is now replaced by  $V_t$  that is free from autocorrelation and we can estimate the regression equation using OLS. OLS, in combination with a variable transformation that results in a corrected error term, is what we call GLS.

#### 2.4.2 GLS when AR(2)

The corresponding transformation in the AR(2) case is very similar. In this case our error term has the following shape:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + V_t \quad (2.27)$$



**THE ANSWER TO  
YOUR LEARNING NEEDS**

**GET QUALITY, FLEXIBLE, AND  
ECONOMICAL TRAINING WHEN  
AND WHERE IT'S NEEDED.**

**SAP Learning Hub**

**SAP**

The advertisement features a man in a dark sweater and glasses holding a tablet, standing in front of a blurred cityscape with a tall building. The SAP logo is in the bottom right corner.

With that in mind we can extend equation (2.26) in the following way:

$$(Y_t - \rho_1 Y_{t-1} - \rho_2 Y_{t-2}) = B_0(1 - \rho_1 - \rho_2) + B_1(X_t - \rho_1 X_{t-1} - \rho_2 X_{t-2}) + V_t \quad (2.28)$$

The whole description above is based on the idea that the autocorrelation coefficient has been known. That is never the case and therefore it must be estimated. The estimated value is often received when you test for autocorrelation. In the Durbin Watson case the test statistic equal

$$DW = 2(1 - \rho) \Rightarrow \rho = 1 - \frac{DW}{2} \quad (2.29)$$

This means that you can use the Durbin Watson test statistic to receive an estimated of the autocorrelation according to (2.29).

In case of higher order of autocorrelation the LM test should be applied. The coefficients in front of the lagged residual terms in (2.21) are estimates of the coefficients in (2.27). Those estimates could therefore be used when transforming the variables according to (2.28).

In the literature you will be able to find more advanced method to estimate the autocorrelation coefficient that could be used when applying GLS. However, statistical software, such as STATA and SPSS, will do most of the job for you. All you have to do is to specify the variables to be used in your model.

### 3. Multicollinearity and diagnostics

Multicollinearity refers to a situation with a high correlation among the explanatory variables within a multiple regression model. For the obvious reason it could never appear in the simple regression model, since it only has one explanatory variable. In chapter 5, book 2 we shortly described the consequences of including the full exhaustive set of dummy variables created from a categorical variable with several categories. We referred to that as to fall in the dummy variable trap. By including the full set of dummy variables, one end up with a perfect linear relation between the set of dummies and the constant term. When that happens we have what is called perfect multicollinearity. In this chapter we will in more detail discuss the issue of multicollinearity and focus on what sometimes is called imperfect multicollinearity which refers to the case where a set of variables are highly correlated but not perfect.

#### *Multicollinearity*

The lack of independence among the explanatory variables in a data set. It is a sample problem and a state of nature that results in relatively large standard errors for the estimated regression coefficients, but not biased estimates.

#### 3.1 Consequences

The consequences of perfect correlation among the explanatory variables is easiest explained by an example. Assume that we would like to estimate the parameters of the following model:

$$Y = B_0 + B_1X_1 + B_2X_2 + U \quad (3.1)$$

where  $X_1$  is assumed to be a linear combination of  $X_2$  in the following way:

$$X_2 = a + bX_1 \quad (3.2)$$

and where  $a$  and  $b$  are two arbitrary constants. If we substitute (3.2) into (3.1) we receive:

$$\begin{aligned} Y &= B_0 + B_1X_1 + B_2(a + bX_1) + U \\ Y &= (B_0 + aB_2) + (B_1 + bB_2)X_1 + U \end{aligned} \quad (3.3)$$

Since (3.1) and (3.2) implies (3.3) we can only receive estimates of  $(B_0 + aB_2)$  and  $(B_1 + bB_2)$ . But since these two expressions contain three unknown parameters there is no way we can receive estimates for all three parameters in (3.1). We simply need more information, which is not available. Hence, with perfect multicollinearity it is impossible to receive an estimate of the intercept and the slope coefficients.

This was an example of the extreme case of perfect multicollinearity, which is not very likely to happen in practice, other than when we end up in a dummy variable trap or a similar situation. More interesting is to investigate the consequences on the parameters and their standard errors when high correlation is present. We will start this discussion with the sample estimator of the slope coefficient  $B_1$  in (3.1) under the assumption that  $X_1$  and  $X_2$  is highly correlated but not perfect. The situation for the sample estimator of  $B_2$  is identical to that of  $B_1$  so it is not necessary to look at both. The sample estimator for  $B_1$  is given by:

$$b_1 = \frac{(r_{Y1} - r_{12}r_{Y2})}{(1 - r_{12}^2)} \frac{S_Y}{S_1} \quad (3.4)$$

The estimator  $b_1$  is a function of  $r_{Y1}$  which is the correlation between  $Y$  and  $X_1$ ,  $r_{12}$  the correlation between  $X_1$  and  $X_2$ ,  $r_{Y2}$  the correlation between  $Y$  and  $X_2$ ,  $S_Y$  and  $S_1$  which are the standard deviations for  $Y$  and  $X_1$  respectively.

The first thing to observe is that  $r_{12}$  appears in both the numerator and the denominator, but that it is squared in the denominator and makes the denominator zero in case of perfect correlation. In case of a strong correlation, the denominator has an increasing effect on the size of the expression but since the correlation coefficient appears in the numerator as well with a negative sign, it is difficult to say how the size of the parameter will change, without any further assumptions. However, it can be shown that the OLS estimators remain unbiased and consistent, which means that estimated coefficients in repeated sampling still will center around the population coefficient. On the other hand, this property says nothing about how the estimator will behave in a specific sample. Therefore we will go through an example in order to shed some light on this issue.

### Example 3.1

Consider the following regression model:

$$Y = B_0 + B_1X_1 + U$$

We would like to know how the estimate of  $B_1$  changes when we include another variable  $X_2$  that is highly correlated with  $X_1$ . Using a random sample of 20 observations we calculate the following statistics.

$$\begin{aligned} S_Y &= 5.1 & r_{Y1} &= 0.843 \\ S_1 &= 5.0 & r_{Y2} &= 0.878 \\ & & r_{12} &= 0.924 \end{aligned}$$

For the simple regression case we receive:

$$b_1 = r_{Y1} \frac{S_Y}{S_1} = 0.843 \times \frac{5.1}{5.0} = 0.86.$$

For the multiple regression case when including both  $X_1$  and  $X_2$  we receive:

$$b_1^* = \frac{0.843 - 0.924 \times 0.878}{1 - 0.924^2} \times \frac{5.1}{5.0} = 0.211.$$

Hence, when including an additional variable the estimated coefficient decreased in size as a result of the correlation between the two variables. Is it possible to find an example where the estimator is increasing in size in absolute terms? Well, consider the case where  $X_2$  is even more correlated with  $X_1$ , let's say that  $r_{12}=0.99$ . That would generate a negative estimate and the small number in the denominator will make the estimate larger in absolute terms. It is also possible to make up an examples where the estimator moves in the other direction. Hence, the estimated slope coefficient could move in any direction as a result of multicollinearity.

In order to analyze how the variance of the parameter estimates change it is informative to look at the equation for the variance. The variance of (3.4) is given by the following expression

$$V(b_1) = \frac{\sum_{i=1}^n e_i^2}{n-3} \frac{1}{(1-r_{12}^2) \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2} \quad (3.5)$$



**MAXIMIZE PRODUCTIVITY**

**HELP YOUR ENTIRE ORGANIZATION BUILD EXPERTISE IN SAP SOFTWARE.**

**SAP Learning Hub**

**SAP**

When the correlation between  $X_1$  and  $X_2$  equals zero, will the variance of the multiple regression coefficient coincide with the variance for the coefficient of the simple regression model. However, when the correlation equals 1 or -1 the variance given by (3.5) will be undefined just as the estimated slope coefficient. In sum, the greater the degree of the multicollinearity, the less precise will be the estimates of the parameters, which means that the estimated coefficients will vary a lot from sample to sample. But make no mistakes; collinearity does not destroy the nice property of minimum variance among linear unbiased estimator. It still has a minimum variance, but minimum variance does not mean that the variance will be small.

It seems like the level of both the estimated parameter and its standard error are affected by multicollinearity. But how will this affect the ratio between them; the t-value. It can be shown that the computed t-value in general will decrease since the standard error is affected more strongly compared to the coefficient. This will usually result in non significant parameter estimates.

Another problem with multicollinearity is that the estimates will be very sensitive to changes in specification. This is a consequence from the fact that there is very little unique variation left to explain the dependent variable since most of the variation is in common between the two explanatory variables. Hence, the parameter estimates are very unstable and sometimes it can even result in wrong signs for the regression coefficient, despite the fact that it is unbiased. A wrong sign is referred to a sign that is unexpected according to the underlying theoretical model, or the prior believes based on common sense. However, sometimes we are dealing with inferior goods which means that we have to be careful with what we call “wrong” sign. Unexpected signs usually require more analysis to understand where it comes from.

### 3.2 Measuring the degree of multicollinearity

Three measures of the degree of multicollinearity are often suggested in the literature: the use of a correlation matrix, the Variance Inflation Factor (VIF), and the tolerance measure. All statistical measures have their limitations, and therefore it is always useful to use several measures when investigation statistical properties of a data set.

Assume that we would like to estimate the parameters of the following model:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + U \quad (3.6)$$

We suspect that the variables are highly correlated and would like to investigate the matter. A natural starting point would be to look at a simple correlation matrix to investigate the pair wise correlations between the variables. That is easily done in any statistical software. Having access to a random sample of 20 observations we received the following results:

**Table 3.1** A correlation matrix for the explanatory variables in (3.6)

	$X_1$	$X_2$	$X_3$
$X_1$	1	0.924	0.458
$X_2$		1	0.085
$X_3$			1

As can be seen from Table 3.1 some of the variables are highly correlated with each other, such as  $X_1$  and  $X_2$ .  $X_1$  is also correlated with  $X_3$  but to a much lower degree and the correlation between  $X_2$  and  $X_3$  is basically zero. From the results of the table we can be sure that that  $B_1$  and  $B_2$  will be difficult to estimate with any good precision, since the correlation between  $X_1$  and  $X_2$  will inflate their standard errors.

To further analyze the multicollinearity we turn to the next measure which is called the **Variance Inflation Factor (VIF)**. It is defined in the following way:

$$VIF(b_i) = \frac{1}{1 - R_i^2} \quad (3.7)$$

where  $R_i^2$  is the squared multiple-correlation coefficient. The squared multiple-correlation coefficient for a specific parameter is a measure of the linear strength between a variable  $X_i$  and the rest of the variables included in the model.

# FAST ADOPTION, FAST ROI

## EQUIP BUSINESS USERS TO ADOPT SAP SOLUTIONS.

SAP Learning Hub, user edition







The squared multiple-correlation coefficient is nothing else than the coefficient of determination received from a auxiliary regression made for each variable against the other variables in the model. That is

$$R_1^2 \text{ is received from: } X_1 = C_{10} + C_{11}X_2 + C_{12}X_3 + U, \quad VIF(b_1) = \frac{1}{1 - R_1^2}$$

$$R_2^2 \text{ is received from: } X_2 = C_{20} + C_{21}X_1 + C_{22}X_3 + U, \quad VIF(b_2) = \frac{1}{1 - R_2^2}$$

$$R_3^2 \text{ is received from: } X_3 = C_{30} + C_{31}X_1 + C_{32}X_2 + U, \quad VIF(b_3) = \frac{1}{1 - R_3^2}$$

When the model contains only two explanatory variables the squared multiple correlation will coincide with the squared bivariate correlation coefficient between the two variables in the model. If you look at (3.5) you will see that the variance inflation factor is included in that expression, and is the factor that is multiplied with the variance of the coefficient of the simple regression model. Hence it is a measure that relates to the case of no correlation, and how the variance is inflated by imposing the correlation. The expression for the variance in the case of more than two explanatory variables has a similar expression but with the squared multiple correlation coefficient instead.

VIF takes values from 1 up to any large number. The closer the multiple-correlation coefficient is to one, the larger the value of VIF. Part of the definition of VIF, is the other multicollinearity statistic called the tolerance. The tolerance measure is the denominator of the VIF expression. Since the square of the multiple-correlation coefficient is a coefficient of determination we could interpret it as such. That means that we have a measure of how large share of the variation in one variable that is explained by a group of other variables. Hence, 1 minus this share would be interpreted as how much of the variation that is left unique for the specific variable and could be used to explain the dependent variable.

When are VIF and the tolerance an indication of a multicollinearity problem? We can shed some lights on that question by an example. Let us go back to model (3.6) and check the VIF and tolerance condition for the variables in that case. Most statistical software has routines for this, and hence you should not need to run the auxiliary regression by your self. Using one such routine in SPSS we received the following regression results and collinearity statistics:

**Table 3.2** Regression results for (3.6)

	P.E.	S.E.	VIF	Tolerance
Constant	117.085	99.782	-	-
$X_1$	4.334	3.016	708.843	0.00141
$X_2$	-2.857	2.582	564.343	0.00177
$X_3$	-2.186	1.595	104.606	0.00956
$R^2$	0.801			
Test of over all significance: F =21.516				
P-value	<0.0001			



From Table 3.2 we see that none of the coefficients are significantly different from zero. We also observe that the coefficient of determination is above 80 percent. That implies that 80 percent of the variation in the dependent variable is explained by the three explanatory variables included in the model. Furthermore, the test of overall significance of the model is highly significant which is in line with the measure of fit. The picture described here is a good example of the consequences of high correlation between the involved variables. It will blow up the standard errors of the model even though the model as such has explanatory power.

When you are exposed to a situation like this, you must go on and calculate some multicollinearity statistics such as VIF and the tolerance for each variable. In the table we see that VIF take large values for all variables if we compare to the case of no correlation that results in  $VIF=1$ . From the analysis of the pair wise correlations we know that the reason for  $X_3$  to have a relatively large VIF number is mainly due to the correlation with  $X_2$ , since the correlation with  $X_1$  is very weak. Even so, VIF for  $X_3$  is 104 and the corresponding tolerance is as low as 0.009 which means that  $X_1$  and  $X_2$  only leave 0.9 percent unique variation left of  $X_3$ 's total variation that can be used in explaining the variation in  $Y$ . The remaining variation was in common with the other two variables and hence must be disregarded. One should observe that even though the pair wise correlations are relatively low, their multiple-regression correlation is much higher, which emphasize the shortcomings of only looking at pair wise correlations.

### 3.3 Remedial measures

With a given specification and data set there is not much one can do about the multicollinearity problem. It could therefore be seen as a state of nature in which data offers no information about some hypothesis that could have been tested using t-tests for the parameters of the model.

Doing nothing is most often not a very attractive alternative. If the alternative of receiving more data is possible, it would be a good solution. It would not solve the multicollinearity problem, but the small unique variation that exist, will be based on more data and if the increase in the number of observations is large enough, it could help increasing the precision of the estimators. To receive more data, and sometimes very much more data, is often very costly and/or time consuming and therefore often not an alternative.

Another alternative would be to change the variable specification. One way of doing that would be to drop one of the variables. If we, in the first place, had an economic relevant specification we know that the estimated parameters will be biased and inconsistent if dropping a relevant variable. Hence, we would only replace one problem with another and this alternative is therefore in general not very attractive.

An alternative approach would be to rethink the model so that it could be expressed in an alternative way. One way of doing that could be to categorize one of the problematic variables. In our example discussed above it was problematic to include  $X_1$  and  $X_2$  in the same regression. But if we replace  $X_1$ , which is a continuous variable, with level indicators (dummy variables) instead it would hedge the strong correlation with  $X_2$  and increase the precision of the estimates. However, that would mean a slightly different model, and we have to be willing to accept that.

In the literature there are other more or less restrictive methods described to handle this problem, and none of them very convincing in their way of reducing the problem. We will therefore not go into any of those more advanced techniques here, since they would require more statistical knowledge that is beyond the scope of this text. But remember, multicollinearity is a state of nature, and is therefore not something that you solve, but instead something that you have to live with.




**JUMP-START CAREERS**

**GIVE STUDENTS ONLINE  
ACCESS TO A VAST BODY  
OF KNOWLEDGE ABOUT  
SAP SOLUTIONS.**

SAP Learning Hub, student edition

**SAP** Learning Hub



## 4. Simultaneous equation models

One important assumption of the basic linear regression model is that the error term has to be uncorrelated with the explanatory variables. If the explanatory variables are in fact correlated with the error term, it would lead to inconsistent estimates of the parameters of the model. In this chapter we will relax this assumption by including additional equations to the model that explains where the correlation is coming from, and discuss the conditions that need to be fulfilled to receive consistent estimate.

### 4.1 Introduction

This chapter will only scratch the surface of the issues involved in estimating simultaneously equations and should therefore be seen as an introduction to the subject. In order for the explanatory variables to be correlated with the error term, they need to be considered random, which was not the case in the previous chapters. The assumption of random explanatory variables does not change anything related to the property of the OLS estimators but it allows for the possibility of being correlated with the error term.

What are the consequences of having the explanatory variable being correlated with the error term? The answer to that question is very similar to the case when we have measurement errors in the explanatory variables, which make the estimates bias and inconsistent. To see this, consider the following simple macro economic model of income determination:

$$Y_t = C_t + I_t \quad (4.1)$$

$$C_t = B_0 + B_1 Y_t + U_t \quad (4.2)$$

with  $Y_t$  being the national income,  $I_t$  investments,  $C_t$  the consumption expenditure and  $U_t$  a stochastic term. Equation (4.1) is an identity and an **equilibrium condition**. Hence this model is formulated under the condition of being in equilibrium, and the equation show how national income is related to consumption and investment in equilibrium. The second, equation given by (4.2), is a **behavioral equation** since it defines the behavior of the consumption expenditure in this economy. Equations with stochastic error terms are to be considered behavioral. Since  $Y_t$  and  $C_t$  are left hand variables in this system of equations, their values are determined by the model. We therefore say that  $Y_t$  and  $C_t$  are **endogenous variables**. We have an additional variable included in the model which is the investment. Since it is a right hand variable in the consumption function we say that it is an **exogenous variable**, which is to say that the value of investments are determined outside the model, it is pre determined. Since investment is determined outside the model it is also uncorrelated with the stochastic term  $U_t$ .

The system of equations can be solved with respect to the two endogenous variables in order to receive their long run expressions. To solve the system with respect to  $Y_t$ , we substitute (4.2) into (4.1) and solve for  $Y_t$ . That results in the following expression:

$$Y_t = \frac{B_0}{1-B_1} + \frac{1}{1-B_1} I_t + \frac{1}{1-B_1} U_t \quad (4.3)$$

In order to receive the long run expression for consumption expenditure, we substitute (4.1) into (4.2) and solve for  $C_t$ :

$$C_t = \frac{B_0}{1-B_1} + \frac{B_1}{1-B_1} I_t + \frac{1}{1-B_1} U_t \quad (4.4)$$

With this setup it is easy to describe the consequences of estimating (4.2) ignoring the fact that it is part of a system. From (4.3) we can see that  $Y_t$  is a function of  $U_t$  which means it is correlated with  $U_t$ . Since  $Y_t$  is correlated with  $U_t$ , we can not use OLS to estimate the coefficients of (4.2) without bias. If consumption expenditure had not been part of this system one could have argued that  $Y_t$  and  $U_t$  in fact are uncorrelated. But when that is not the case we see from (4.3) how they are related.

It should now be obvious that the OLS estimators are biased in small samples due to the correlation between  $Y_t$  and  $U_t$ . But are they also inconsistent? That is, if we increase the number of observations to a very large number, will the estimators still be biased? To see this consider the OLS estimator for  $B_1$ :

$$b_1 = \frac{\sum_{t=1}^T (Y_t - \bar{Y}) C_t}{\sum_{t=1}^T (Y_t - \bar{Y})^2} = B_1 + \frac{\sum_{t=1}^T (Y_t - \bar{Y}) U_t}{\sum_{t=1}^T (Y_t - \bar{Y})^2} \quad (4.5)$$

This expression was developed in chapter 3, book 1 (see (3.12)). If we take the expected value of the estimator we will receive:

$$E[b_1] = B_1 + E \left[ \frac{\sum_{t=1}^T (Y_t - \bar{Y}) U_t}{\sum_{t=1}^T (Y_t - \bar{Y})^2} \right] \quad (4.6)$$

The problem with the expectation on the right hand side is that  $Y_t$  is a random variable and correlated with  $U_t$ , and for that reason we can not proceed as in chapter 3, book 1. Furthermore, since the expectation is a linear operator we have that  $E[A/B] \neq E[A]/E[B]$ , which further complicates the problem. Even though this makes it clear that the estimator no longer is unbiased, we do not know how the second component on the right hand side of (4.6) behave in large samples. It can be shown that the limit of the OLS estimator is given by the following expression:

$$\lim_{t \rightarrow \infty} b_1 = B_1 + \frac{(1-B_1)\sigma_U^2}{\sigma_I^2 + \sigma_U^2} \quad (4.7)$$

(4.7) show that in the limit the sample estimator still deviate from the population parameter, which means that the bias remains in large samples.

Correlation between the error term and the explanatory variables in a single equation model using OLS would lead to:

- Biased and inconsistent parameter estimates
- Invalid tests of hypothesis
- Biased and inconsistent forecasts

## 4.2 The structural and reduced form equation

From the previous discussion we learned that when an equation belongs to a system of equation, estimating them separately using OLS would lead to biased and inconsistent estimates. Hence, in order to be able to estimate the parameters of the equation system it is important to consider the whole system since they interact with each other. Before going into issues of estimation we need to define some more concepts. Consider the following system of equations:



**LEARN BY DOING**

**DEVELOP EXPERTISE  
IN SAP SOLUTIONS  
THROUGH EXPLORATION  
AND PRACTICE.**

SAP Live Access

**SAP** Learning Hub



$$C_t = A_0 + A_1 Y_t + U_{1t} \quad (4.8)$$

$$I_t = B_0 + B_1 R_t + U_{2t} \quad (4.9)$$

$$Y_t = C_t + I_t + G_t \quad (4.10)$$

It is a macro economic model that extends the example from the previous section and is based on three equations. It is an income determination model, with two behavioral equations; one for consumption expenditure  $C_t$ , and one for net-investments  $I_t$ . The consumption function is a function of income  $Y_t$  and the investment function is a function of interest rate  $R_t$ . The income equation that specifies the equilibrium condition is a function of consumption, investment and government spending  $G_t$ . This model has three endogenous variables,  $C_t$ ,  $I_t$ , and  $Y_t$ , and two exogenous variables  $R_t$  and  $G_t$  that are pre determined.

The system of equations given by (4.8)-(4.10) describes the structure of the economy that we would like to investigate. For that reason these equations are called **structural equations**. The coefficients of the structural equations represent the direct effect of a change in one of the explanatory variables. If we take (4.9) as an example,  $B_1$  represents the marginal propensity to invest as a result from a change in the interest rate. This represents the **direct effect** of a change in interest rate on the net-investment.

Assume that we increase the interest rate. That will have a direct effect on the investments in this model, which in a second step via the equilibrium condition will have an effect on the income. The income in its term will affect the consumption level, and since income is endogenous it will have an effect the error term  $U_1$  since they are correlated. The initial change in the interest rate, will in this way, affect the components in the system until the effect reaches its equilibrium level.

We can therefore talk about two types of effect; the short run effect and the long run effect. The long run effect can be received from the long run relationship that can be determined by solving the structural equation system with respect to the endogenous variables. To solve the system for  $Y_t$  we simply substitute (4.8) and (4.9) into (4.10) and solve for  $Y_t$ . If we do that we receive:

$$Y_t = \frac{A_0 + B_0}{1 - A_1} + \frac{B_1}{1 - A_1} R_t + \frac{1}{1 - A_1} G_t + \frac{U_{1t} + U_{2t}}{1 - A_1} \quad (4.11)$$

If we do the similar thing with respect to the other two endogenous variables we would receive the following expressions:

$$C_t = \frac{A_0 + A_1 B_0}{1 - A_1} + \frac{A_1 B_1}{1 - A_1} R_t + \frac{A_1}{1 - A_1} G_t + \frac{A_1 U_{2t} + U_{1t}}{1 - A_1} \quad (4.12)$$

$$I_t = B_0 + B_1 R_t + U_{2t} \quad (4.13)$$



By solving the structural system of equations with respect to the endogenous variables we have determined the **reduced form equations** for income, consumption and investment. The coefficients of the reduced form equations represent the full effect when the system is in equilibrium. The full effect of a change in interest rate on income is represented by  $B_1/(1-A_1)$ . It is also called the **interest rate multiplier** on income. There is a corresponding multiplier related to consumption and investments that can be found in their reduced form equations. Observe that the reduced form equation for investments only is a function of interest rate. Government spending does not have any affect on the investment, even though it has an effect on consumption and income.

The nice thing with the reduced form equations is that they may be estimated separately using OLS. That is, the coefficients in the reduced form equations can be consistently estimated using OLS. Since the structural parameters are part of the reduced form coefficients it is sometimes possible to indirect find the structural coefficient using the estimated values or the reduced form coefficients. For that to be possible, certain requirements need to be fulfilled. The structural coefficients must be exactly identified.

### 4.3 Identification

In order to be able to estimate the structural equation coefficients they need to be identified. So, what do we mean by that? To give an intuitive feeling for its meaning we will give an example before going into any formal and mechanical tests.

Consider the following two equation system:

$$Q = A_0 + A_1P + A_2X_1 + U_1 \quad (\text{Supply}) \quad (4.14)$$

$$Q = B_0 + B_1P + U_2 \quad (\text{Demand}) \quad (4.15)$$

This system contains two endogenous variables ( $P$  and  $Q$ ), and one exogenous ( $X_1$ ) variable. These two equations represent a demand and supply system for a given market. The question is if any of these two equations are identified. That is to ask if the parameters of the two equations can be estimated consistently.

It turns out that the demand function is identified while the supply function is not. To see this, consider Figure 4.1. In order to identify the demand function we need some exogenous variation that could help us trace out the function. That could be done using the supply function. The supply function contains an exogenous variable  $X_1$  and the supply function takes a new position for each value of  $X_1$ . In that process we identify the demand function. But in the demand function we have nothing unique that does not appear in the supply function so it is impossible to move the demand function while holding the supply function fixed. Hence it is the presence of an exogenous variable in one equation that allows us to estimate the parameters of the other equation. If  $X_1$  had been included in both equations there would have been no unique variation in any of the equations and hence no equation had been identified. However, if another exogenous variable,  $X_2$ , had been introduced and placed in the demand function, we would receive some exogenous variation that could help us to identify the supply function. In that case both equations would have been identified.

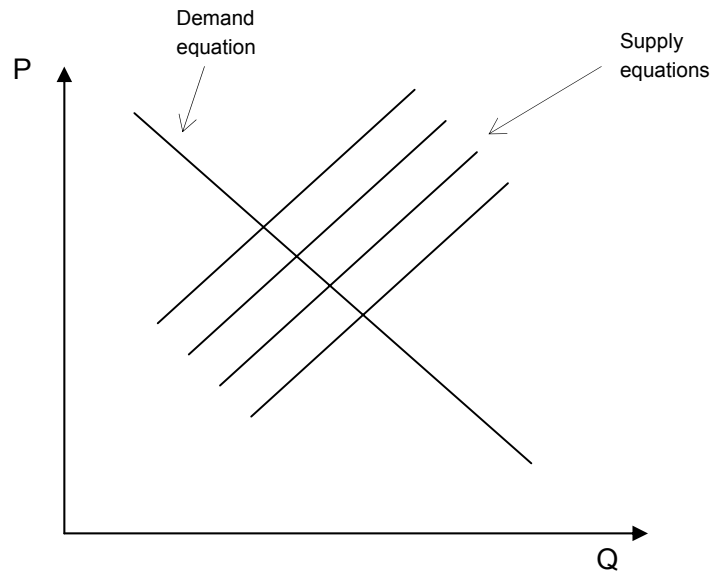


Figure 4.1 Demand and supply system

The process of identifying equations can be formalized in a decision rule that specifies the conditions that have to be fulfilled in order to identify one or several equations in a system. In the literature two rules are described and one is slightly easier to use than the other.

**HANDS-ON PRACTICE  
FOR EFFECTIVE LEARNING**

**EXPERIENCE SAP  
SOFTWARE FIRSTHAND  
TO BUILD KNOWLEDGE  
AND ENHANCE SKILLS.**

SAP Live Access

**SAP** Learning Hub

**SAP**



### 4.3.1 The order condition of identification

The first decision rule for identification is the so called order condition. This rule specifies the necessary conditions for identification and is the more popular one of the two rules that will be discussed.

Unfortunately it is not a sufficient rule, which means that it is possible that the equation is undefined even though the order condition says it is identified. However, in a system with only two equations, the order condition will work well and can be trusted.

Define the following variables:

$M =$  The number of endogenous variables in the model  
 $K =$  The number of variables (endogenous and exogenous) in the model excluded from the equation under consideration.

**The order condition states that:**

- 1) If  $K = M - 1 \Rightarrow$  The equation is exactly identified
- 2) If  $K > M - 1 \Rightarrow$  The equation is over identified
- 3) If  $K < M - 1 \Rightarrow$  The equation is under identified

When checking the order condition you have to do it for each equation in the system.

#### Example 4.1

Consider the following system:

$$Y_1 = A_0 + A_1 Y_2 + A_2 X_1 + U_1 \quad (4.16)$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + U_2 \quad (4.17)$$

Use the order condition to check if the equations are identified. In order to do that, we need to determine the value of  $M$  and  $K$ . This system contains two endogenous variables and the total number of variables, endogenous as well as exogenous, is 4.

For the first equation we have  $M-1=1$  and  $K=1$  since  $X_2$  is excluded from (4.16). Since  $M-1=K$  we have that the first equation is exactly identified.

For the second equation we have  $M-1=1$  and  $K=1$  since  $X_1$  is excluded from (4.17). Since  $M-1=K$  we have that also the second equation is exactly identified. When all the equations of the model are identified we say that the model is identified since we are able to estimate all the structural parameters.

**Example 4.2**

Consider the following system:

$$Y_1 = A_0 + A_1 Y_2 + A_2 X_1 + A_3 X_3 + U_1 \quad (4.18)$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + U_2 \quad (4.19)$$

In this example we have two endogenous variables and three exogenous variables with a total of five variables.  $M-1$  will in this example equal 1 as before since we still have only two endogenous variables. Will the equations be identified in this case? The first equation contains four variables which means that one variable has been excluded from the equation, that is,  $X_2$  does not appear in equation 1 and  $K=1$ . Since  $M-1=K$  the equation is exactly identified.

The second equation includes three variables which mean that two variables have been excluded. That is,  $X_1$  and  $X_3$  are not included in equation 2. That means that  $K=2$ , which means that  $M-1 < K$  which leads to the conclusion that equation 2 is over identified.

**4.3.2 The rank condition of identification**

The rank condition is slightly more complicated when dealing with larger systems of equations, but when using only two equations it is as easy as the order condition. The rank condition is a necessary and sufficient condition, which means that if we can identify the equations using the rank condition we can be sure that the equation really is identified. The rank condition investigates whether two or more equations are linearly dependent on each other, which would be the case if the sum of two equations would equal a third equation in the model. If that is the case it is impossible to identify all structural parameters. The basic steps in this decision rule is best described by an example.

**Example 4.3**

Consider the following system of equations:

$$Y_1 = A_0 + A_1 Y_3 + A_2 X_1 + A_3 X_3 + U_1 \quad (4.20)$$

$$Y_2 = B_0 + B_1 Y_1 + B_2 X_2 + B_3 X_3 + U_2 \quad (4.21)$$

$$Y_3 = C_0 + C_1 Y_2 + A_2 X_1 + A_3 X_2 + U_3 \quad (4.22)$$

This system contains three endogenous variables ( $Y_1, Y_2, Y_3$ ) and three exogenous variables ( $X_1, X_2, X_3$ ), which means that we in total has six variables. The first step in checking the rank condition is to put up a matrix that for each equation mark which of the six variables that are included (marked with 1) and which that are excluded (marked with 0) from the equation. For our system we receive the following matrix:

*Matrix for the rank condition*

	$Y_1$	$Y_2$	$Y_3$	$X_1$	$X_2$	$X_3$
Equation 1	1	0	1	1	0	1
Equation 2	1	1	0	0	1	1
Equation 3	0	1	1	1	1	0

In order to check the rank condition for the first equation we have to proceed as follows: Delete the first row and collect the columns for those variables of the first equation that were marked with zero. For equation 1,  $Y_2$  and  $X_2$  was marked with zero, and if we collect those two columns we receive:

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

If this matrix contains less than  $M-1$  rows or columns where all elements are zero, equation 1 will not be identified.  $M$  refers to the number of equation just as in the order condition, which means that  $M-1=2$ . Since we have two rows and two columns and none of them contains only zeros we conclude that equation 1 is identified.

For equation 2 we proceed in the same way. We delete the second row and collect those columns where the elements of the second row were marked with a zero. For equation 2 that was the case for  $Y_3$  and  $X_1$ , which is to say that these two variables was not included in equation 2. The resulting matrix for this case then becomes:



**ANYTIME, ANYWHERE**

**LEARNING ABOUT  
SAP SOFTWARE HAS  
NEVER BEEN EASIER.**

SAP Learning Hub – the choice of  
when, where, and what to learn

**SAP** Learning Hub

**SAP**

$$\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}$$

It looks in the same way as for equation 1, which means that we have two rows and two columns that is not only zeros. The same procedure should be done for the third equation and if you do that you will see that it is identified as well.

When using larger systems it is quite possible that the order condition says that a particular equation is identified even though the rank condition says it is not. When that happens it might still be possible to generate estimates, but those estimates will not have any economic meaning since they will represent averages of those equations that are linear combinations of each other. Hence, you should not be content that you have received identified results just because the order condition says so and the econometric software generates results for you. When using systems of more than two equations you should also confirm the identification using the rank condition.

## 4.4 Estimation methods

Once we have confirmed that our model is identified we can proceed with the estimation of the parameters of the structural coefficients. In this section we will present two methods of estimation that can be used to estimate coefficients of a simultaneous equation system.

### 4.4.1 Indirect Least Squares (ILS)

When all the equations are exactly identified one can use the method of Indirect Least Square to estimate the coefficients of the structural equations. It is done by the following three steps:

- 1) Form the reduced form equations
- 2) Estimate the coefficients of the reduced form using OLS
- 3) Use the estimated coefficients of the reduced form to derive the structural coefficients.

#### **Example 4.4 (ILS)**

Consider the following simple macro economic model:

$$Y_t = C_t + I_t \quad (4.23)$$

$$C_t = B_0 + B_1 Y_t + U_t \quad (4.24)$$

This model has two endogenous variables ( $Y_t$  and  $C_t$ ) and one exogenous variable ( $I_t$ ), and we would like to estimate the coefficients of the behavioral equation. Since one of the variables of the model is excluded from the consumption function it is identified according to the order condition. The two structural equations could be used to form the reduced form equations for consumption. If we do that we receive:

$$C_t = \pi_0 + \pi_1 I_t + V_t \quad (4.25)$$

(13.3) and (13.4) show how the reduced form coefficients are related to the structural coefficients. By using the estimated values of the reduced form coefficients we can solve for the structural coefficients. We have:

$$\pi_0 = \frac{B_0}{1 - B_1} \quad (4.26)$$

$$\pi_1 = \frac{B_1}{1 - B_1} \quad (4.27)$$

(4.26) and (4.27) can now be used to solve for  $B_0$  and  $B_1$ . Since (4.27) is an equation with only one unknown we solve for  $B_1$  first (remember that  $\pi_1$  is an estimate and therefore a number in this expression). Once we receive the value of  $B_1$  we can use it in (4.26) to solve for  $B_0$ . Hence we receive:

$$B_0 = \frac{\pi_0}{1 - \pi_1} \quad \text{and} \quad B_1 = \frac{\pi_1}{1 - \pi_1}$$

In order to determine the standard errors for  $B_0$  and  $B_1$  we can use linear approximations to their expression based on the standard errors and covariance of the reduced form estimated coefficients. It can be shown that the corresponding variance for  $B_0$  and  $B_1$  is:

$$V(B_0) \approx a^2 \sigma_0^2 + b^2 \sigma_1^2 + 2ab \sigma_{01}$$

$$V(B_1) \approx b^2 \sigma_1^2$$

with  $a = \frac{1}{1 - \pi_0}$  and  $b = \frac{1}{(1 - \pi_1)^2}$  and where  $\sigma_0^2$  is the variance of  $\pi_0$ ,  $\sigma_1^2$  the variance for  $\pi_1$  and  $\sigma_{12}$  the covariance between  $\pi_0$  and  $\pi_1$ .

ILS will result in consistent estimates but will still be biased in small samples. When using larger systems with more variables and equations it is often burdensome to find the estimates, and in those cases the equations are often over identified, which means that ILS cannot be used. For that reason ILS is not used very often in practice. Instead a much more popular method called 2SLS is used.

#### 4.4.2 Two Stage Least Squares (2SLS)

The procedure of 2SLS is a method that allows you to receive consistent estimates of the structural coefficient when the equations are exactly identified as well as over identified. However, the estimates will still be biased in small samples.

Consider the following model

$$Y_1 = A_0 + A_1Y_2 + A_3X_1 + A_4X_2 + U_1 \quad (4.28)$$

$$Y_2 = B_0 + B_1Y_1 + B_3X_3 + B_4X_4 + U_2 \quad (4.29)$$

This model has two endogenous variables and four exogenous variables. The first equation (4.28) contains four variables which means that from a total of six variables, two has been omitted. That means that it is over identified. The same can be said about the second equation (4.29), which means that the model is identified. Since both equations are over identified we cannot estimate the structural parameters using ILS, but instead we are forced to use 2SLS. We will now focus the discussion on the estimation of the first equation.



The basic steps of 2SLS applied for equation (4.28):

- Step 1 Derive the reduced form equation for  $Y_2$  and estimate the predicted value of  $Y_2$  ( $\hat{Y}_2$ ) on the reduced form using OLS.

$$\hat{Y}_2 = \hat{\pi}_0 + \hat{\pi}_1 X_1 + \hat{\pi}_2 X_2 + \hat{\pi}_3 X_3 + \hat{\pi}_4 X_4$$

- Step 2 Replace  $Y_2$  in equation (4.28) with its predicted value from the reduced form and estimate the coefficient of the model using OLS.

$$Y_1 = A_0 + A_1 \hat{Y}_2 + A_3 X_1 + A_4 X_2 + U_1$$

If these two steps are applied we will receive consistent estimates of the parameters in (4.28). That is, since we replace the endogenous variable with its predicted value, it is no longer correlated with the residual term. Hence, the problem is solved. Remember that  $Y_2 = \hat{Y}_2 + V_2$  which implies that the stochastic variable  $Y_2$ , consist of two parts, one that is a linear combination of the exogenous (predetermined) variables and one random part. A group of exogenous variables are by necessity uncorrelated with the random term.

Observe that  $X_1$  and  $X_2$  both appear in the specification of  $Y_1$  and  $\hat{Y}_2$ , which means that there will be a correlation between the explanatory variables  $X_1$  and  $X_2$  and  $\hat{Y}_2$ . This correlation will not be perfect unless  $X_3$  and  $X_4$  also is included in the structural model of the first equation and is therefore nothing to worry about. But if that happens, the equation would not pass the order condition for identification.

There is one additional complication to be aware of when working with 2SLS. When the predicted value is included in the specification, the variance of the error term will not be correct. To see this we will consider a simplified version of a model to make it clear where the problem appear. Consider the following equation:

$$Y_{1i} = B_1 Y_{2i} + U_{1i} \quad (4.30)$$

In order to receive consistent estimates of  $B_1$  we replace  $Y_2$  with its predicted value and estimate the parameters of following regression model using OLS:

$$Y_{1i} = B_1 \hat{Y}_{2i} + (U_{1i} + B_1 U_{2i}) = B_1 \hat{Y}_{2i} + V_1 \quad (4.31)$$

The estimator of the slope coefficient is therefore given by the following expression:

$$b_1 = \frac{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2) Y_{1i}}{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2)^2} = \frac{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2) (B_1 Y_{2i} + U_{1i})}{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2)^2} = B_1 + \frac{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2) U_{1i}}{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2)^2} \quad (4.32)$$

We have concluded that this form of the estimator is consistent but not unbiased. Since it is consistent we need to compare it with its asymptotic variance, that is, the formula of the variance when the number of observation is very large (has gone to infinity). It can be shown that the asymptotic variance of this sample estimator is given by the following expression:

$$V(b_1) = \frac{\sigma_U^2}{\sum_{i=1}^n (\hat{Y}_{2i} - \bar{\hat{Y}}_2)^2} \quad (4.33)$$

**THE ANSWER TO  
YOUR LEARNING NEEDS**

**GET QUALITY, FLEXIBLE, AND  
ECONOMICAL TRAINING WHEN  
AND WHERE IT'S NEEDED.**

**SAP** Learning Hub





This is good, because it is very similar to the variance given by the standard OLS. So what is the problem? The problem is related to the estimated variance of the error term. When running the regression using (4.31) our estimated residual would be given by:

$$\hat{\sigma}_V^2 = \frac{1}{n} \sum_{i=1}^n (Y_{1i} - b_1 \hat{Y}_{2i})^2 \quad (4.34)$$

Whereas the estimated residual should be given by the following expression

$$\hat{\sigma}_U^2 = \frac{1}{n} \sum_{i=1}^n (Y_{1i} - b_1 Y_{2i})^2 \quad (4.34)$$

(4.34) is based on the observed variable  $Y_2$  multiplied with the sample estimator  $b_1$  given by (4.32), rather than the predicted version of the variable. Hence in order to receive consistent estimates of the standard errors, one has to use (4.34). When using commercial software with routines for 2SLS they automatically make the correction. But if you run 2SLS in two steps, as described above, you need to correct the standard errors, before you can perform and hypothesis testing.

***In sum, the 2SLS has the following properties:***

- It generates biased but consistent estimates
- The distribution of the estimators are normally distributed only in large samples
- The variance is biased but consistent when using (4.34)

## A. Statistical tables

Table A1

Area below the standard normal distribution:  $P(Z \leq z)$ 

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999
3.1	0.99903	0.99906	0.9991	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.9994	0.99942	0.99944	0.99946	0.99948	0.9995
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.9996	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.9997	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

**Example:** Find the probability that  $Z$  is less than 1.33:  $P(Z \leq 1.33) = 0.90824$

Table A2

Right tail critical values for the t-distribution

Degrees of freedom	Probability						
	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.050	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000001	3.077685	6.313749	12.70615	31.82096	63.6559	318.2888
2	0.816497	1.885619	2.919987	4.302656	6.964547	9.924988	22.32846
3	0.764892	1.637745	2.353363	3.182449	4.540707	5.840848	10.21428
4	0.740697	1.533206	2.131846	2.776451	3.746936	4.60408	7.17293
5	0.726687	1.475885	2.015049	2.570578	3.36493	4.032117	5.893526
6	0.717558	1.439755	1.943181	2.446914	3.142668	3.707428	5.207548
7	0.711142	1.414924	1.894578	2.364623	2.997949	3.499481	4.785252
8	0.706386	1.396816	1.859548	2.306006	2.896468	3.355381	4.500762
9	0.702722	1.383029	1.833114	2.262159	2.821434	3.249843	4.29689
10	0.699812	1.372184	1.812462	2.228139	2.763772	3.169262	4.143658
11	0.697445	1.36343	1.795884	2.200986	2.718079	3.105815	4.024769
12	0.695483	1.356218	1.782287	2.178813	2.68099	3.054538	3.929599
13	0.69383	1.350172	1.770932	2.160368	2.650304	3.012283	3.852037
14	0.692417	1.345031	1.761309	2.144789	2.624492	2.976849	3.787427
15	0.691197	1.340605	1.753051	2.131451	2.602483	2.946726	3.732857
16	0.690133	1.336757	1.745884	2.119905	2.583492	2.920788	3.686146
17	0.689195	1.333379	1.739606	2.109819	2.56694	2.898232	3.645764
18	0.688364	1.330391	1.734063	2.100924	2.552379	2.878442	3.610476
19	0.687621	1.327728	1.729131	2.093025	2.539482	2.860943	3.579335
20	0.686954	1.325341	1.724718	2.085962	2.527977	2.845336	3.551831
21	0.686352	1.323187	1.720744	2.079614	2.517645	2.831366	3.527093
22	0.685805	1.321237	1.717144	2.073875	2.508323	2.818761	3.504974
23	0.685307	1.319461	1.71387	2.068655	2.499874	2.807337	3.484965
24	0.68485	1.317835	1.710882	2.063898	2.492161	2.796951	3.466776
25	0.68443	1.316346	1.70814	2.059537	2.485103	2.787438	3.450186
26	0.684043	1.314972	1.705616	2.055531	2.478628	2.778725	3.43498
27	0.683685	1.313704	1.703288	2.051829	2.472661	2.770685	3.42101
28	0.683353	1.312526	1.70113	2.048409	2.467141	2.763263	3.408204
29	0.683044	1.311435	1.699127	2.045231	2.46202	2.756387	3.396271
30	0.682755	1.310416	1.69726	2.04227	2.457264	2.749985	3.385212
35	0.681564	1.306212	1.689573	2.03011	2.437719	2.723809	3.340028
40	0.680673	1.303076	1.683852	2.021075	2.423258	2.704455	3.306923
45	0.679981	1.30065	1.679427	2.014103	2.412116	2.689594	3.281457
50	0.679428	1.298713	1.675905	2.00856	2.403267	2.677789	3.261375
60	0.678601	1.295821	1.670649	2.000297	2.390116	2.660272	3.231689
80	0.677569	1.292224	1.664125	1.990065	2.373872	2.638699	3.195237
100	0.676951	1.290075	1.660235	1.983972	2.364213	2.625893	3.173773
∞	0.67449	1.281551	1.644853	1.959966	2.326351	2.575835	3.090245

Note: The smaller value at the head of each column is the area in one tail, the larger value is the area in both tails.

Table A3

Right tail critical value of the Chi-Square distribution

Degrees of freedom	Probability				
	0.100	0.050	0.025	0.010	0.005
1	2.705541	3.841455	5.023903	6.634891	7.8794
2	4.605176	5.991476	7.377779	9.210351	10.59653
3	6.251394	7.814725	9.348404	11.34488	12.83807
4	7.779434	9.487728	11.14326	13.2767	14.86017
5	9.236349	11.07048	12.83249	15.08632	16.74965
6	10.64464	12.59158	14.44935	16.81187	18.54751
7	12.01703	14.06713	16.01277	18.47532	20.27774
8	13.36156	15.50731	17.53454	20.09016	21.95486
9	14.68366	16.91896	19.02278	21.66605	23.58927
10	15.98717	18.30703	20.4832	23.20929	25.18805
11	17.27501	19.67515	21.92002	24.72502	26.75686
12	18.54934	21.02606	23.33666	26.21696	28.29966
13	19.81193	22.36203	24.73558	27.68818	29.81932
14	21.06414	23.68478	26.11893	29.14116	31.31943
15	22.30712	24.9958	27.48836	30.57795	32.80149
16	23.54182	26.29622	28.84532	31.99986	34.26705
17	24.76903	27.5871	30.19098	33.40872	35.71838
18	25.98942	28.86932	31.52641	34.80524	37.15639
19	27.20356	30.14351	32.85234	36.19077	38.58212
20	28.41197	31.41042	34.16958	37.56627	39.99686
21	29.61509	32.67056	35.47886	38.93223	41.40094
22	30.81329	33.92446	36.78068	40.28945	42.79566
23	32.00689	35.17246	38.07561	41.63833	44.18139
24	33.19624	36.41503	39.36406	42.97978	45.55836
25	34.38158	37.65249	40.6465	44.31401	46.92797
26	35.56316	38.88513	41.92314	45.64164	48.28978
27	36.74123	40.11327	43.19452	46.96284	49.64504
28	37.91591	41.33715	44.46079	48.27817	50.99356
29	39.08748	42.55695	45.72228	49.58783	52.3355
30	40.25602	43.77295	46.97922	50.89218	53.67187
35	46.05877	49.80183	53.20331	57.34199	60.27459
40	51.80504	55.75849	59.34168	63.69077	66.76605
45	57.50529	61.65622	65.41013	69.9569	73.16604
50	63.16711	67.50481	71.42019	76.1538	79.48984
60	74.397	79.08195	83.29771	88.37943	91.95181
80	96.5782	101.8795	106.6285	112.3288	116.3209
100	118.498	124.3421	129.5613	135.8069	140.1697

Table A4

Right tail critical for the F-distribution: 5 percent level

$n/m$	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	$\infty$
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95	248.02	249.26	250.10	251.14	252.20	253.25	254.32
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32	1.22	1.00

Note:  $m$  = degrees of freedom for the numerator $n$  = degrees of freedom for the denominator