

FYI Handout #5: More on the Error Term

As we have seen, the linear regression model

$$(1) \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

traditionally assumes that the errors ε_i are independent and normally distributed with a mean of zero and a standard deviation of σ .

These basic assumptions of the linear regression model can be relaxed. For example, the assumption of normality of the error term can be dropped if the sample size is large. Because of the central limit theorem, the sampling distributions of the estimates of β_j 's (that is the least-squares coefficients b_0, b_1, \dots, b_k) are approximately normal, so the inference procedures based on the t statistic are still valid.

The assumption that the standard deviation of the error term is constant (homoskedasticity) can also be relaxed. Instead it can be assumed that the error term is heteroskedastic, meaning that its standard error varies from observation to observation. In that case the standard deviation of ε_i is denoted by σ_i . This situation is more common in cross-sectional than in time series data. An example could be a regression of savings on income for a cross-section of individuals. Low-income individuals do not have a lot of financial capacity for saving, because they need to use most of their income for basic expenditures in housing, clothing, meals, and transportation. On the other hand, high-income people can choose between spending most of their income in luxurious goods and services and saving little or spending more moderately and saving a high proportion of their income. In this case the standard deviation of the error term will be higher for higher values of the explanatory variable (income).

Among the several tests for heteroskedasticity, a popular one is White's test. It consists of running an auxiliary regression of the squared residual of the original regression, $e_i^2 = (y_i - \hat{y}_i)^2$ on the explanatory variables of the original regression, plus their squares and cross-products. For illustration, if the original regression has two explanatory variables, the auxiliary regression looks like

$$(2) \quad e_i^2 = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{1i}^2 + \gamma_4 x_{2i}^2 + \gamma_5 x_{1i} x_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

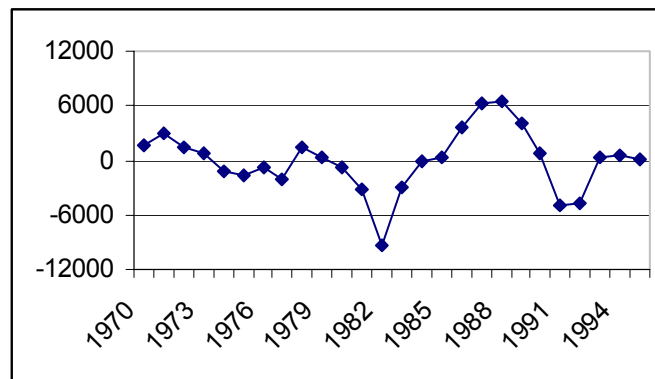
Sometimes, the auxiliary regression does not include the cross-products. The rationale of the test is as follows. If the error term is in fact homoskedastic, the residuals of the regression should **not** be associated with the explanatory variables; therefore the R^2 of the auxiliary regression should be low. The test statistic suggested by White to test the null hypothesis $H_0 : \sigma_i = \sigma$ for all i (homoskedasticity) against the two-sided alternative $H_0 : \sigma_i \neq \sigma$ for at least one i (heteroskedasticity) is the product between the sample size n

and the R^2 of the auxiliary regression. This statistic is distributed as a chi-square with k degrees of freedom, where k is the number of explanatory variables in the original regression (2 in the example).

White's heteroskedasticity test is a standard option in Eviews. After estimating your regression, click on View in the Equation window and then select Residual Tests. The program gives you the option of performing White's test with and without cross-products in the auxiliary regression. After selecting one of this options, the program displays the value of the test statistic $Obs \cdot R\text{-squared}$ followed by its P -value. In addition, the program shows you estimation details of the auxiliary regression.

What do you do if the White test rejects the null hypothesis of homoskedasticity (say if the P -value of the test statistic is less than 5%)? The main problem with heteroskedasticity is that it affects the accuracy of the standard errors of the regression estimates (though not the estimates themselves). Fortunately, White proposed a method to correct the standard errors, which is incorporated as an option in Eviews. To produce White-corrected standard errors select Options in the Equation Specification window and then check the box next to "Heteroskedasticity Consistent Covariance". The printout with the regression results will have a note saying that you have used this correction.

Another of the basic assumptions of the linear regression model that sometimes is not satisfied by the data is the errors ε_i are independent. Particular in the case of time series data, the error term may be correlated across subsequent observations. When this happens it is said that the error term is autocorrelated. An example of an autocorrelated error term is depicted in the following graph, which shows the residuals of a regression of aggregate consumption on GDP for Australia with data between 1970 and 1995.



Clearly, the residuals do not seem to follow a random pattern around zero, as we should expect if the error terms were independent. The residuals tend to be positive for a few years, then negative for another few years and so on. The question is why this happens?

More often than not, the finding of autocorrelation in the residuals of a time series regression indicates a problem of specification of the model. For example, an important

variable could have been omitted. The simple consumption model from where the residuals in the graph were obtained, omits important factors in the consumption decision, such as expectations about the future. Then, it is generally good practice, when one finds a problem of autocorrelation to think hard what is causing it before jumping to “fix it”. Moreover, autocorrelation may be completely unrelated to any economic explanation because it often is caused by the way the data was constructed. For example, high frequency data (such as daily data) are often reported as monthly averages by statistical agencies. This averaging may result in seeming autocorrelation when this data is used in regression analysis. This further support the advise of not running to “fix the problem” until you have not thoroughly explored its possible causes.

Supposing that the autocorrelation is legitimate, a second problem is that it can take different forms. A common assumption is that the error term at time t is related to the error term at time $t - 1$:¹

$$(3) \quad \varepsilon_t = \rho\varepsilon_{t-1} + v_t ,$$

where ρ (known as the coefficient of autocorrelation) can take values between -1 and 1 and v_t are independent or “white noise” errors.

As in the case of heteroskedasticity, autocorrelation tests are based on the regression residuals. The most conventional test is the Durbin-Watson test, which is based on the following statistic:

$$(4) \quad d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}, \quad t = 1, 2, \dots, T$$

The numerator of this statistic can be written as

$$(5) \quad \sum_{t=2}^T (e_t - e_{t-1})^2 = \sum_{t=2}^T e_t^2 + \sum_{t=2}^T e_{t-1}^2 - 2\sum_{t=2}^T e_t e_{t-1} .$$

If the error term is **not** autocorrelated ($\rho = 0$ in Equation 3), the last term in Equation (5) should be close to zero. Then, because each of the first two terms are similar to the denominator in Equation (4), the value of the statistic should be close to 2. Next consider the case where the error term is strongly positively autocorrelated (ρ close to 1). In that case the last term of the numerator is approximately equal to minus the sum of the first two terms, implying that the d statistic is close to zero. Finally, if the error term is strongly negatively autocorrelated (ρ close to -1), the last term of the numerator is similar to the sum of the first two, implying that d is close to 4.

¹ This is known as an autoregressive error of order one or AR(1).

In economics time series, the most common problem is positive autocorrelation, so you should expect to find a value of the d statistic that is less than 2. If this value is low (say around 1 or less), you should check carefully the model for specification errors and the data. If you are convinced that the autocorrelation is “legitimate” and that there is a good economic reason to account for it, you can then “fix” the problem. A simple way to do it is as follows.

Suppose that you want to estimate the following regression model with an autocorrelated error:

$$(6) \quad y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad t = 1, 2, \dots, T$$

$$(7) \quad \varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad -1 \leq \rho \leq 1, \quad t = 1, 2, \dots, T$$

where the errors v_t are independent with zero mean and standard error σ . If we lag Equation (6) one period and multiply by ρ on both sides, we get

$$(8) \quad \rho y_{t-1} = \rho \beta_0 + \rho \beta_1 x_{t-1} + \rho \varepsilon_{t-1}.$$

Subtracting Equation (8) from Equation (6) we obtain

$$(9) \quad y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + \varepsilon_t - \rho \varepsilon_{t-1}, \text{ or}$$

$$(10) \quad y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + v_t$$

This transformed equation has an independent error term; therefore it can be estimated without problems. The traditional econometric literature provides many methods to estimate ρ , in order to implement Equation (10). A relatively simple approach is to write Equation (10) as

$$(11) \quad y_t = (1 - \rho)\beta_0 + \rho y_{t-1} + \beta_1 x_t - \beta_1 \rho x_{t-1} + v_t.$$

That is to say, a lagged dependent variable y_{t-1} and the lagged values of the explanatory variables (in this case, just one: x_{t-1}) are added to the explanatory variables. This is very simple to implement in Eviews. For example if your original regression has Y as dependent variable and X as explanatory variable, you write in the Equation Specification box **Y C X** (C stands for the constant). In order to add the lagged variables, write **Y(-1) X(-1)**, where the variable names followed by (-1) are lagged values. After you have estimated Equation (11), you can interpret the estimated coefficient of the lagged dependent variable y_{t-1} as an estimate of the coefficient of autocorrelation of the error term.