**Serial Correlation in Regression Analysis**

One of the assumptions of both simple and multiple regression analysis is that the error terms are independent from one another – they are uncorrelated. The magnitude of the error at some observation, *i* (*i = Yi – A – BXi*) has no effect on the magnitude or sign of the error at some other observation, *j.* This assumption is formally expressed as *E(**ij*) = 0 for all *i ≠* j, which means that the expected value of all pair-wise products of error terms is zero. If indeed, the error terms are uncorrelated, the positive products will cancel those that are negative leaving an expected value of 0.

If this assumption is violated, although the estimated regression model can still be of some value for prediction, its usefulness is greatly compromised. The estimated regression parameters, *a*, *b*1, *b*2, . . . ,*bk*, remain unbiased estimators of the corresponding true values, *A*, *B*1, *B*2, . . ,*Bk*, leaving the estimated model appropriate for establishing point estimates of *A, B*, etc., and the model can be used for predicting values of *Y* for any given set of *X* values. However, the standard errors of the estimates of the regression parameters, *e.g., sb* are significantly underestimated which leads to erroneously inflated *t* values. Because testing hypotheses about the slope coefficients (*i.e.,* *Bk*) and computing the corresponding confidence intervals rely on the calculated *t* values as the test statistics, the presence of correlated error terms means that these types of inferences cannot be made reliably.

Although there are many ways this assumption might be violated, the most common occurs with time series data in the form of *serial correlation*. In the context of time series, the error in a period influences the error in a subsequent period – the next period or beyond*.*  Consider this: if there are factors (other than the independent variables) making the observation at some point in time larger than expected, (*i.e.,* a positive error), then it is reasonable to expect that the effects of those same factors linger creating an upward (positive) bias in the error term of a subsequent period. This phenomenon, for obvious reasons, is called *positive* *first-order serial correlation* – by far the most common manner in which the assumption of independence of errors is violated.[[1]](#footnote-1)

Of course, it’s easy to imagine situations where the error term in one period does not affect the error in the next period (no first order serial correlation), but biases the error terms two periods later (second-order serial correlation). Consider a time series influenced by quarterly seasonal factors: a model that ignores the seasonal factors will have correlated error terms with a lag of four periods. We will concern ourselves, however, only with first order serial correlation.

The Durbin Watson test is a well known formal method of testing if serial correlation is a serious problem undermining the model’s inferential suitability (*e.g.,* assessing the confidence in the predicted value of a dependent variable). The test statistic of the Durbin-Watson procedure is *d* and is calculated as follows:



Recall that *et* represents the observed error term (*i.e.,* residuals) or (*Yt –Ŷt*) = *Yt – a – bXt*. It can be shown that the value of *d* will be between zero and four; zero corresponding to perfect positive correlation and four to perfect negative correlation. If the error terms, *et*and *et-*1, are uncorrelated, the expected value of *d* is 2. The further *d* is below 2 the stronger the evidence for the existence of positive first order serial correlation and *vice versa*. Unfortunately, the Durbin-Watson test can be inconclusive. The critical values of *d* for a given level of significance, sample size and number of independent variables are tabulated as pairs of values: *DL* and *DU* (a table is provided in “Course Documents/Statistical Tables folder). If the test statistic, *d* falls between these two values the test is inconclusive. The formal test of positive first order serial correlation is as follows:

Ho **(no serial correlation)

H1 **(positive serial correlation)

If *d* < *DL*reject Ho, while if *d* > *DU* do not reject Ho.

Although negative first order serial correlation is far less likely, the statistic *d* can be used to test for the existence of negative serial correlation as well. For this test the critical limits are 4 - *DL* and 4 - *DU*. The test then is:

Ho **(no serial correlation)

H1 *<*(negative serial correlation)

If *d* < 4 - *DU* do not reject Ho, if *d* > 4 - *DL*reject Ho.

These decision points can be summarized as in the following figure.

*DL*

*DU*

4*-DU*

4 -*DL*

0

> 0

= 0

< 0

?



**Example:**

The model we used earlier, *Y = A + BX* + ** where *Y* is the *sales revenue* and *X* is the *advertising budget*, is actually a time series; the successive pairs of *Y* and *X* values are observed monthly. Therefore, there is concern that first-order serial correlation may exist. To test for positive serial correlation we calculate the error terms, *ei*. Note that the Excel regression tool will generate the error terms upon request by checking the residuals box in the regression tool dialog. The errors (residuals) are given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RESIDUAL OUTPUT | |  |  |  |  |
|  |  |  |  |  |
| *Observation* | *Predicted Yi* | *Residuals*  *(et)* | *ei-1* | *(ei – ei-1)2* |
| 1 | 1473.967 | 348.4015 |  |  |
| 2 | 1388.173 | 11.23608 | 348.4015 | 113680.5 |
| 3 | 1242.198 | -41.7989 | 11.23608 | 2812.713 |
| 4 | 1154.249 | -431.179 | -41.7989 | 151616.8 |
| 5 | 2453.899 | 133.2528 | -431.179 | 318583.1 |
| 6 | 1518.662 | -23.3486 | 133.2528 | 24523.97 |
| 7 | 733.0114 | 16.43126 | -23.3486 | 1582.434 |
| 8 | 1291.79 | 224.1726 | 16.43126 | 43156.46 |
| 9 | 908.6782 | 288.5593 | 224.1726 | 4145.651 |
| 10 | 1639.618 | -650.965 | 288.5593 | 882705.1 |
| 11 | 1357.892 | -0.07709 | -650.965 | 423654.5 |
| 12 | 1855.751 | 125.3145 | -0.07709 | 15723.06 |

Notice that the denominator of the Durbin Watson statistic is SSE = = 900,722.1. To calculate the numerator we shift residuals down by one row and calculate the sum of the squared differences starting with the second row. These calculations are done by appending the last two columns to the table. The sum of the last column which gives the numerator is 1982184. We can now compute the Durbin-Watson statistic, *d* as 1982184/900,722.1 = 2.200661.

Formally Ho **(no serial correlation)

H1 **(positive serial correlation)

For *n* =12, *k* = 1 and *DL* = .97 and *DU*= 1.32. Since *d* > *DU* we accept the null hypothesis that there is no significant positive serial correlation. You may wish to formally test the existence of negative serial correlation.

1. Negative first-order serial correlation occurs when a positive (negative) error is followed by a negative (positive) error. That is, the current error negatively influences the following error. [↑](#footnote-ref-1)