Serial correlation (also called “autocorrelation”) is said to exist when the error terms of any pair of observations are correlated with one another. In other words, the error terms are not independent of one another and thus $E(\varepsilon_i \varepsilon_i) \neq 0$ (i.e., the covariance is not equal to zero). Serial correlation tends to exist primarily in time series data – data where the same phenomena (GDP, stock market indices, etc.) is measured in repeated time periods – and panel data (where a cross-section of a population is surveyed more than once). In both cases, the same person or thing is measured in two time periods; the time periods are treated as different observations. Time series and panel data tend to have this problem because the value of the dependent variable is primarily explained by what the value of the dependent was in the previous period. For instance, the value of the Dow Jones today is primarily explained by the value of the Dow Jones yesterday. Thus, if the error term was large yesterday it is very likely to be large today – thus $E(\varepsilon_i \varepsilon_i) \neq 0$. Pindyck and Rubinfeld provide two other explanations for serial correlation. First, errors may become correlated because the same measurement error is made again and again. Second – and more likely – is that omitted variable bias accumulates over multiple observations of the same unit. For the purposes of this class, we will only deal with so-called first order serial correlation – also known as AR(1) – which occurs when an observation at $t+1$ is correlated with an observation at $t$.

When serial correlation exists, one of the Gauss-Markov Conditions has failed. In this case, we can say for certain that the failure leads to inefficient estimates – the standard error is consistently underestimated. We will tend to reject the null hypothesis that $\beta = 0$ too readily. On average, serial correlation does not lead to bias or inefficiency in our coefficient estimates. However, in practice, OLS regression in the presence of serial correlation tends to generate misleading, if technically unbiased, estimates. The reason for this is related to the nature of time series data. For instance, if we wish to study GDP’s relationship to the Dow Jones, we only have reliable data on the last 50 to 60 years. Often, we need data on 100 or 200 periods (in this case, years) to accurately understand a trend. OLS fits a line very well to data we have; it cannot fit a line to data we don’t have – yet those missing points may be very influential in determining the “true” relationship.

In the example below, I have regressed the Dow Jones Industrial Average (DJI) against U.S. GDP (in 1996 dollars). One regression was for the (actual) data from 1990 to 2002. The second regression used the 1990-2002 data and hypothetical (i.e., made-up) data for 2003-2013. Let’s pretend that I’m God (I love these promotions) and know this to be the “true” path of the stock market and that the period 1990-2013 is sufficient to establish the correct relationship between the Dow Jones and GDP. Mere mortals may only measure outcomes from 1990 to 2002. You can see that Stata fitted very closely a line through the 1990-2002 data; it also found a tight fit through the 1990-2013 – but these lines are quite different. Mere mortals simply lack the data needed to estimate the relationship well. Serial correlation makes the problem worse; the “dot-com” bubble of the late 1990s is a perfect example. From year to year, the dot-com mania drove the stock market further and further away from its historical path – the “unusually” large random factors in 1998, 1999, and 2000 were correlated with one another and added up to a line that
showed a much steeper relationship. To find it correctly, we need to regress on the both “inflation” and “deflation” of the bubble, but our data only covers the “inflation” periods. The correlated errors exacerbated the problem.¹

However, if we could purge the serial correlation between values in the 1990-2002 data, our line would fit the phenomena much better – we would find the “underlying” trends more clearly. Thus, though OLS is unbiased on average, the practical impact of serial correlation is to badly mis-state the coefficient values and most certainly to overstate a) the confidence we have in the coefficient estimates and b) the “goodness of fit” (R²) of the model.

**Finding Serial Correlation in Data**

In practice, it is not always possible to tell whether there is serial correlation in a time series or panel data set – or if there is “enough” to affect our levels of significance or coefficient estimates. However, it is a problem for which you must always test. Most researchers assume serial correlation exists in time series or panel data unless evidence to the contrary is presented. The most widely used test for serial correlation is the Durbin-Watson (DW) statistic. The DW stat measures the ratio of the sum of squared difference in pairs of residuals to the sum of squared residuals for each observation alone. The formula for the statistic is not particularly important – Stata will calculate it for you. What is important is understanding the complex interpretation of it. Let’s look at an example, drawn from a data set on gross state product (GSP).

To use any of the “post-regression commands” for time series issues like serial correlation, one must first tell Stata what variable in the dataset contains the time index – the variable that tells when two

¹ To be very clear: there are two issues in the data I presented above. The first is a problem of insufficient data – lacking data on a full boom-bust cycle. The second is that the correlated errors tended to make our estimated relationship even less realistic.
observations are in the same time period. In this data set, the time index variable is “year” – it contains the year as a number (1969, 1970, etc.). The command to “declare” this variable as the time index is “tsset”:

```
.tsset year
  time variable:  year, 1969 to 1986
```

We then proceed to run our regression (don’t worry about the function form now – we’re only interested in the serial correlation properties of this data):

```
.reg  lngsp lnemp lnkpri lnpub
```

```
Source |       SS       df       MS              Number of obs =      18
-------------+------------------------------           F(  3,    14) =   50.63
Model |  .061339833     3  .020446611           Prob > F      =  0.0000
Residual |   .00565433    14  .000403881           R-squared     =  0.9156
-------------+------------------------------           Adj R-squared =  0.8975
Total |  .066994163    17  .003940833           Root MSE      =   .0201

------------------------------------------------------------------------------
  lngsp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lnemp |   .8285251   .2508164     3.30   0.005     .2905774    1.366473
  lnkpri |   .4229073   .1136382     3.72   0.002     .1791776     .666637
  lnpub |  -.2033786   .1896595    -1.07   0.302    -.6101576    .2034005
  _cons |  -3.579139   3.734449    -0.96   0.354    -11.58873    4.430458
------------------------------------------------------------------------------
```

The post-regression command is “dwstat” in Stata 8 and before. In Stata 9 the command is estat dwatson.

```
.dwstat
Durbin-Watson d-statistic(  4,    18) =  .9543101
```

Stata does not offer a significance level for the DW stat; it only reports the value of the statistics itself – in this case .9543. To interpret the coefficient, you will need to use the table at the back of Pindyck and Rubinfeld (page 610). There are two parameters that matter regarding the DW stat: the number of variables in the regression (k) and the number of observations (N). For reasons known only to the folks in Texas, Stata does not report the number of variables correctly – it includes the constant when it should not. So for this example, we should be looking up k=3, N=18, even though Stata reported the statistics as “(4,18).”

If you take a look at the table on page 610, you will see that two numbers are reported – d_L and d_U. These are the so-called “critical values.” For reasons explained in Pindyck and Rubinfeld, the DW stat has four different regions that have a different meaning – they stretch along a number line between 0 and 4.

<table>
<thead>
<tr>
<th>Positive serial corr</th>
<th>Indeterminate</th>
<th>No serial correlation</th>
<th>Indeterminate</th>
<th>Negative serial corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>d_L</td>
<td>d_u</td>
<td>2</td>
<td>4-d_u</td>
</tr>
</tbody>
</table>

Values between 0 and d_L and between 4- d_L are indicative of serial correlation (at the 95% confidence level). Value between d_u and 4- d_u are indicative of no serial correlation (at the 95% confidence level). Values between d_L and d_u and between 4- d_u and 4- d_L are indeterminate – they may or may not be indicative of serial correlation (the statistics do not reach the 95% confidence level in this areas).

In our example, the table on page 610 for k = 3 and N = 18, has entries of d_L = .93 and d_u = 1.69. So, for this data and model, our zones are:
Our DW stat of 0.9543 falls into the “indeterminate” zone – but just barely. We are quite close to a score that would find serial correlation with a 95% certainty. We should at least try the standard correction to see if it improves the DW stat.

Correcting for Serial Correlation

There are several corrections for serial correlation. The most heavily used correction is the Cochrane-Orcutt procedure. This procedure uses a two step procedure that tries to a) estimate the correlation $\rho$ between pairs of observations and b) uses generalized differencing to make use of this estimated correlation. Generalized differencing tends to remove the serial correlation – see Pindyck and Rubinfeld for an in-depth explanation. In Stata, we use the `prais` command with the “corc” option included at the end to estimate a regression using the Cochrane-Orcutt procedure. Before using `prais` you must declare which variable contains the year identifier using `tsset`. Below, a `prais` was used to correct for serial correlation in our GSP data. Note that the estimated $\rho$ between pairs of observations is 0.6522 – the residuals seem to be highly correlated with one another.

```
. prais lngsp lnemp lnkpri lnkpub, corc
Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.5229
Iteration 2:  rho = 0.6077
Iteration 3:  rho = 0.6345
Iteration 4:  rho = 0.6445
Iteration 5:  rho = 0.6487
Iteration 6:  rho = 0.6506
Iteration 7:  rho = 0.6515
Iteration 8:  rho = 0.6519
Iteration 9:  rho = 0.6521
Iteration 10: rho = 0.6522
Iteration 11: rho = 0.6522
Iteration 12: rho = 0.6523
Iteration 13: rho = 0.6523
Iteration 14: rho = 0.6523
Cochrane-Orcutt AR(1) regression -- iterated estimates
```

```
Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  3,    13) =   17.95
Model |   .01664714     3  .005549047           Prob > F      =  0.0001
Residual |  .004017805    13  .000309062           R-squared     =  0.8056
-------------+------------------------------           Adj R-squared =  0.7607
Total |  .020664944    16  .001291559           Root MSE      =  .01758

------------------------------------------------------------------------------
lngsp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnemp |   1.079805   .2986772     3.62   0.003     .4345519    1.725058
lnkpri |    .300484   .1876143     1.60   0.133    -.1048322    .7058001
lnkpub |   .0447519   .5159484     0.09   0.932    -1.069887    1.159391
_cons  |  -8.691443   6.265587    -1.39   0.189    -22.22742    4.844535
-------------+----------------------------------------------------------------
rho |   .6522747
------------------------------------------------------------------------------
```

Durbin-Watson statistic (original)  0.954310
Durbin-Watson statistic (transformed)  1.408649
At the end of the Cochrane-Orcutt output Stata reports the DW stat for the original estimation and for the "transformed" estimate (i.e., after doing Cochrane-Orcutt). Unfortunately, our results are still indeterminate, though we are now closer to the critical value for “no serial correlation” than we were at first. Note also how substantially our estimates have changed – the coefficient on $K_{pub}$ has gone from negative to positive. Our levels of statistical significance have also changed, and the $R^2$ has also fallen. Nonetheless, we cannot be certain that all effects of serial correlation have been purged from this regression; only a DW stat between 1.69 and 2.31 would indicate that. To do better, we would need more data.