Definition
Serial correlation occurs when error terms are correlated across observations.

In its most general sense this means that instead of the expected correlation between error terms of two observations being zero

$$E(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

we instead see some correlation between observations so that

$$E(\varepsilon_i, \varepsilon_j) \neq 0 \quad (i \neq j)$$

which is a violation of one of the classical assumptions.

Pure Serial Correlation

Now, this correlation might occur on the basis of just about anything, but the most important seems to be correlations over time in time series data. For example, if you’re modeling unemployment rates, it may be that the error term associated with one month’s observation might be correlated with the previous month’s error term. In other words, if last month’s unemployment rate was unusually high then this month’s unemployment rate will probably be unusually high, too.

The most common form of serial correlation is first order serial correlation, which states that the error term in one period is correlated with the error term in the preceding period, or the previous observation of the error term

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (-1 < \rho < 1)$$

where

$\rho$ is the correlation coefficient between the two errors (also called, in this case, the first-order autocorrelation coefficient)

$u_t$ is a nonserially correlated error term
If $\rho > 0$ then there is positive serial correlation, meaning that unusually high values of the dependent variable tend to occur together and that unusually low values of the dependent variable tend to occur together. That is, you tend to see long runs of high or low values.

If $\rho < 0$ then there is negative serial correlation, meaning that an unusually high value is likely to be followed by an unusually low value of the dependent variable. That is, values tend to fluctuate back and forth from period to period.

There may be other forms of autocorrelation. Observations of a variable could be seasonally or quarterly correlated or might be correlated from the observation from the same month from one year ago:

$$
\varepsilon_t = \rho \varepsilon_{t-4} + u_t \quad (-1 < \rho < 1)
$$

$$
\varepsilon_t = \rho \varepsilon_{t-12} + u_t \quad (-1 < \rho < 1)
$$

There might also be higher order forms of autocorrelation. That is, the error associated with an observation might depend on the error from the two previous observations:

$$
\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t \quad (-1 < \rho_1 < 1, -1 < \rho_2 < 1)
$$

Impure Serial Correlation

Impure serial correlation occurs because of an omitted variable. If a model is estimated that excludes an independent variable that should be included, then the resulting residuals could appear serially correlated, even if the error terms from the proper model wouldn’t be.

In terms of equations, this looks like:

Proper Model: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$

Incorrect Model: $Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon^*_t$

Error Term from Incorrect Model: $\varepsilon^*_t = f(\beta_2 X_{2t} + \varepsilon_t)$

This will be a problem particularly when $X_{2t}$, the excluded variable, is serially correlated. This might also be a problem if $X_{2t}$ changes in a steady way over time.
Consequences of Serial Correlation

1. Coefficient estimates are not biased by pure serial correlation. This isn’t so much a consequence as it is a lack of a consequence. There’s no bias introduced by pure serial correlation, so the expected values of the coefficient estimates are equal to the true values of the coefficients:

\[ E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad E(\hat{\beta}_2) = \beta_2 \]

1A. Impure serial correlation, because it is based on excluded variables, may have some bias associated with the estimated coefficients as a result of there being excluded variables.

2. The variance of the estimated coefficients will rise, so the OLS estimates are not necessarily the best estimates of the coefficients.

3. Serial correlation makes the standard errors of the coefficients larger in a way that OLS can’t detect, so estimated coefficients will appear to be more significant than they really are and the model as a whole will appear to have more predictive power than it really does.

In other words, the estimated t-stats will be too big and the significance values of the estimated coefficients will be too small. In addition, the R^2 values will be too big, the F-stats will be too big and the p-values associated with the F-statistics will be too small. So your model will look better than it really is.

Detection of Serial Correlation

One way to try and detect serial correlation is to simply estimate the model of your choice (it is based on good underlying theory, isn’t it?) and save the residuals. Then plot the residuals against your time variable (year, month, day or whatever) and see if there tend to be long runs of positive or negative values that would suggest some degree of positive correlation. This process has the particularly nasty-sounding name runs test.

The most commonly used test for serial correlation is the Durbin-Watson d Test.

This test is valid under the following circumstances:
1. The regression model includes a constant.
2. The serial correlation is first order serial correlation: \( \varepsilon_t = \rho \varepsilon_{t-1} + u_t \)
3. There are no lagged explanatory variables.

The equation for the Durbin-Watson d statistic is:
If there is extreme positive serial correlation, then \( d=0 \).

If there is extreme negative serial correlation, then \( d \approx 4 \).

If there is no serial correlation, then \( d \approx 2 \).

Because the situation of negative serial correlation is usually difficult to explain theoretically, there isn’t usually concern about a particularly large value of the Durbin-Watson statistic.

**Remedies for Serial Correlation**

We’ll address pure serial correlation in a minute, but the first thing to consider when thinking about addressing serial correlation is the impure variety, which is caused by excluded independent variables. If you seem to have some serial correlation, the first and most fundamental thing to think about is whether you have excluded some important explanatory variable and whether this could be responsible for your serial correlation.

After you have dispensed with the impure variety, then you can try some fancy moves to try to deal with pure serial correlation.

**Generalized Least Squares (GLS)**

GLS is a method of dealing with pure first-order serial correlation. Here’s the basic idea.

Start with an equation that has a problem with first order autocorrelation:

\[
Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t
\]

but because there is autocorrelation, the error term is really

\[
\varepsilon_t = \rho \varepsilon_{t-1} + u_t
\]
where $\rho$ is the coefficient of serial correlation and $u_t$ is a proper, uncorrelated error term. This gives us:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \rho e_{t-1} + u_t$$

What we'd really like to do is estimate the equation with only the $u_t$ terms in it and none of those pesky, serially correlated $\varepsilon_t$ terms.

To try and do this, we might note that:

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \varepsilon_{t-1}$$

$$\rho Y_{t-1} = \rho \beta_0 + \rho \beta_1 X_{1t-1} + \rho \varepsilon_{t-1}$$

and a little subtraction gives us

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_{1t} - \rho X_{1t-1}) + u_t$$

The problem is that you'll never know what the serial correlation coefficient $\rho$ really is, but it can be estimated and the regression can be done based on the estimate of $\rho$.

Gosh, how do I estimate $\rho$?

One way to estimate $\rho$ is to calculate the Durbin-Watson statistic and then calculate $\rho$ based on that according to the formula:

$$\hat{\rho} \approx 1 - \frac{d}{2}$$

A second way is to use the Cochrane-Orcutt iterative method. To do this, estimate the OLS equation you like in spite of it suffering from serial correlation, then regress the residuals on the lagged residuals with no constant term:

$$e_t = \rho e_{t-1} + u_t$$

What you will get from this regression is an estimate of $\rho$, $\hat{\rho}$. You then use this estimated value to generate values for variables in the equation:

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_{1t} - \rho X_{1t-1}) + u_t$$
using $\hat{\rho}$ instead of $\rho$.

You then repeat these steps (re-do the regression of lagged residuals) over and over again until your estimates of $\rho$ stop changing. This will give you a good estimate of the value of $\rho$ and the value of the slope coefficient $\beta_1$.

Keep a Trigger Lock on GLS
GLS is a tool, and like any tool it should be used with caution.

1. GLS won’t work if you have impure serial correlation. Emphasize the fundamentals and think about the specification of your model before you go whipping out the GLS.

2. Serial correlation might not be such a big deal, depending on what you’re trying to do with the model. Serial correlation doesn’t bias the coefficient estimated, so if all you’re looking for is estimates of slope coefficients, serial correlation isn’t necessarily a problem.

**Conclusion**
Serial correlation occurs when the error terms of a model are correlated over observations. This usually occurs with models based on time series data.

Serial correlation can be caused by a model specification that excludes an important explanatory variable.

Serial correlation doesn’t bias coefficient estimates, but can make things seem significant when they really aren’t.

You can detect serial correlation through runs tests or through a test based on the Durbin-Watson statistic.

You can correct for serial correlation by using generalized least squares, but this will only improve your results if you have pure serial correlation.