

Regression Diagnostics

Issues of Independence

Independence means that the errors associated with one observation are not correlated with the errors of any other observation. Of course, we want the assumption of independence to hold in our regression. In other words, we want that, once all regression covariates are considered, there are no further correlations – that is, dependence – between measures. There can be dependence in several situations. The main possibilities can be summarized in two different scenarios: 1) temporal dataset, and 2) hierarchical (or multilevel) dataset. Both situations are related to an omission-bias problem that you are not able to deal with your model (and that, therefore, produces the violation of the Issue of Independence).

1) Temporal dataset

A first way in which the assumption of independence can be violated is when data are collected on the same variables over time. Let's say that we collect inflation data every semester for 12 years. In this situation, it is likely that the errors for observations estimated in two adjacent semesters will be more highly correlated than for observations that are more distant in time. This is known as autocorrelation. This feature is typical of time-series data: observations falling close to each other in time are not independent but rather tend to be correlated with each other. If inflation is low today, it will be probably low in the next semester as well. And so on. This MAY produce also correlated error (OR MAYBE NOT...). Note that we always look for a lack of independence in the errors once discounted for all the PREDICTORS, i.e., it is not enough that the value of our DV of today is similar to its value in the past to produce autocorrelation. Substantively, autocorrelation is equivalent to tricking yourself into believing that you have more information than you really do.

We will focus here on the way to detect the most common problem in most temporal data – that is, we will control if the error term in our linear regression model follows an AR(1) process (first-order serial correlation in the errors).

What does it mean more formally that the error term in a linear regression model follows an AR(1) process? For the linear model $Y_t = \beta_1 X_t + \mu_t$ the AR(1) process can be written as: $\mu_t = \alpha \mu_{t-1} + \epsilon_t$.

How to check for it? First way: we could employ the Durbin–Watson test. The Durbin–Watson test can be applied only when the regressors are strictly exogenous. A regressor x is strictly exogenous if $\text{Corr}(x; u_t) = 0$ for all x and t (i.e. exogenous in the past, in the present and in the future). This

implies that the Durbin–Watson statistic cannot be used with models where lagged values of the dependent variable (i.e., Y_{t-1}) are included as regressors.

The null hypothesis of the test is that there is no first-order autocorrelation. The Durbin–Watson d statistic can take on values between 0 and 4, and under the null d is equal to 2. Values of d less than 2 suggest positive autocorrelation ($\alpha > 0$), whereas values of d greater than 2 suggest negative autocorrelation ($\alpha < 0$).

Calculating the exact distribution of the d statistic is difficult, but empirical upper and lower bounds have been established based on the sample size and the number of regressors. Extended tables for the d statistic can be found here (or anywhere else through Google): https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf

To test for positive autocorrelation at significance α , the test statistic d is compared to lower and upper critical values (dL, α and dU, α):

If $d < dL, \alpha$, there is statistical evidence that the error terms are positively autocorrelated.

If $d > dU, \alpha$, there is no statistical evidence that the error terms are positively autocorrelated.

If $dL, \alpha < d < dU, \alpha$, the test is inconclusive.

Positive serial correlation is serial correlation in which a positive error for one observation increases the chances of a positive error for another observation.

If the observed test statistic value is greater than 2, then you want to test the null hypothesis against the alternative hypothesis of negative first-order autocorrelation. To test for negative autocorrelation at significance α , the test statistic $(4 - d)$ is compared to lower and upper critical values (dL, α and dU, α):

If $(4 - d) < dL, \alpha$, there is statistical evidence that the error terms are negatively autocorrelated.

If $(4 - d) > dU, \alpha$, there is no statistical evidence that the error terms are negatively autocorrelated.

If $dL, \alpha < (4 - d) < dU, \alpha$, the test is inconclusive.

Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another.

An example: using Klein's (1950) data on the US economy, we first fit an OLS regression of consumption on the government wage bill (minimum salary).

Let's use the dataset consumption.dta (hint: Stata provides access to several example datasets):

We use the `tsset` command to declare data to be time-series data, with year (`yr`) as the variable indicating time:

```
tsset yr
```

The output indicates that the time variable ranges from year 1920 to year 1941, with an interval that is equal to 1 unit.

We now regress consumption on minimum salary:

```
regress consump wagegovt
```

If we assume that `wagegovt` is a strictly exogenous variable, we can use the Durbin–Watson test to check for first-order serial correlation in the errors.

```
estat dwatson
```

Durbin-Watson d -statistic(2, 22) = .3217998

The Durbin–Watson d -statistic, 0.32, is far from the center of its distribution ($d = 2.0$). Given 22 observations and two regressors (including the constant term) in the model, the lower 5% bound is about 1.23949, much greater than the computed d statistic. We can therefore reject the null of no first-order serial correlation.

If we are not willing to assume that `wagegovt` is strictly exogenous, we could instead use Durbin’s alternative test or the Breusch–Godfrey test for first-order serial correlation. Because we have only 22 observations, we will use the small option.

```
estat durbinalt, small
```

```
estat bgodfrey, small
```

If we are willing to assume that the errors follow an AR(1) process and that `wagegovt` is strictly exogenous, we could refit the model trying to correct for it using a “newey model” (note that this is really a rough way to deal with such problem. Example:

```
newey consump wagegovt, lag(1)
```

`newey` produces Newey–West standard errors for coefficients estimated by OLS regression. As you remember, the Huber/White/sandwich robust variance estimator (see White 1980) produces consistent standard errors for OLS regression coefficient estimates in the presence of heteroskedasticity. The Newey–West (1987) variance estimator is an extension that produces consistent estimates when there is autocorrelation in addition to possible heteroskedasticity. The Newey–West variance estimator handles autocorrelation up to and including a lag of m , where m is specified by stipulating the `lag()` option. Thus, it assumes that any autocorrelation at lags greater than m can be ignored. If `lag(0)` is specified, the variance estimates produced by `newey` are simply the Huber/White/sandwich robust variances estimates calculated by `regress, vce(robust)`.

Alternatively, you can try to differentiate your dependent variable or you can use other models – i.e., time-series regression models that allow to deal with the error process explicitly).

```
gen delta_con = d.consump
```

```
regress delta_con wagegovt
```

```
estat dwatson
```

```
gen delta2_con = d2.consump
```

```
regress delta2_con wagegovt
```

```
estat dwatson
```

Remember: Having a temporal dataset is just a necessary BUT NOT sufficient condition to have autocorrelation in your residuals. It all depends on the model you are estimating!!!

Let us run the following model:

```
regress consump wagegovt wagepriv  
estat dwatson  
estat durbinalt, small  
estat bgodfrey, small
```

Here we do not have problems with AR(1) anymore