

## 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](https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf)

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

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  $\alpha$ , the test statistic  $(4 - d)$  is compared to lower and upper critical values ( $dL, \alpha$  and  $dU, \alpha$ ):

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!!!

HINT: if you run the following model you do not have anymore problems with AR(1):

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

## 2) Hierarchical (or multilevel) dataset

Many research designs in the social sciences have a **hierarchical structure**. Such hierarchies are produced because the population is hierarchically structured – that is, **observations tend to be grouped into higher-level units**.

Consider the case of collecting data from students in ten different elementary schools. Probably, students within the same school will tend to be more like one another than students from different schools. In other words, their errors are not independent, once controlled for all your independent variables. Or suppose that you want to test the relationship between satisfaction with democracy in European countries and life satisfaction. Probably, the respondents within each country will tend to be more like one another than respondents from different countries. Once again, their errors are not independent, **after controlling** for your potential predictors of satisfaction with democracy.

As already discussed, OLS regression assumes that the residuals are independent. But if they are not, you will get biased standard errors!

Let's open the satisfaction for democracy dataset (dataset satisfaction with democracy.dta). The dataset contains data on about 30,000 respondents coming from 29 different countries.

We pretend that `demo_satisf` (i.e., satisfaction with democracy) is an interval-level variable even if it is actually an ordinal variable (this is just to run a simple example!).

It is very well possible that citizens' levels of satisfaction for democracy within each country are not independent, and this could lead to residuals that are not independent within countries. **How to deal with violation of independence when the data are hierarchical?**

### Option a) Using the Cluster option

### Option b) Fixed effects regression

### Option c) Random effects regression

## Option a) Using the Cluster Option

We can use the `cluster` option to indicate that the observations in our dataset are clustered into higher-level groups (here, countries) and that the observations may be correlated within each group (here, country) but would be independent between groups (here, countries).

Now, we can run regress with the `cluster` option (where `id` is the variable identifying each single country). We do not need to include the `robust` option since `robust` is implied with `cluster`.

```
reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use
```

```
reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use, r
```

```
reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use, cluster(id)
```

Note that the standard errors have changed substantially, much more so, than the change caused by the `robust` option itself. Look for example at the coefficient for `radio_use`!

As with the `robust` option, when we use the `cluster` option the estimates of the coefficients are the same as the OLS estimates, but now the standard errors take into account that the observations within each country are non-independent. These standard errors are computed based on aggregate scores for the 29 countries, since the national levels of satisfaction for democracy should be independent from each other! If you have a very small number of clusters compared to your overall sample size, it is possible that the standard errors are quite larger than with the OLS regression. For example, if there were only 3 countries, the standard errors would be computed on the aggregate scores for just 3 countries. So **NEVER use cluster standard errors when you have a low number of clusters (i.e., lower than 20 – or probably better – lower than 40!)**.

Note that when you are using the `cluster` option you treat the existence of correlation within countries as a nuisance/problem that you want to avoid. In other models (e.g. options: b; c) you model precisely this correlation.

## Option b) Fixed effects regression

A possible alternative to cluster standard errors is to employ so called “fixed effects”. Fixed effects regression is a method for controlling for omitted variables in the dataset when the omitted variables vary across groups or entities (i.e., European countries in our previous example) but do not change across observations within a given group or entity (i.e., do not change across Italian citizens, do not change across French citizens, and so on). The fixed effects regression model has  $n$  different intercepts, one for each entity. These intercepts can be represented by a set of binary (or indicator) variables. These binary variables absorb the influences of all omitted variables that differ from one entity to the next but are constant over observations within those entities.

Consider the following regression model where we want to explain `demo_satisf` of individual  $i$  living in country  $j$  ( $Y_{ij}$ ) with `life_satisf` (a variable measuring satisfaction with life for individual  $i$  in country  $j$ ) ( $X_{ij}$ ):

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + e_{ij} \quad (1)$$

Where  $Z_j$  is an unobserved variable that varies from one country to the other but does not change by definition over the respondents living in that same country (for example,  $Z_j$  represents unmeasured national cultural attitudes toward `demo_satisf`). Because  $Z_j$  varies from one country to the next one, but is constant over respondents within the same country, the regression model in (1) can be interpreted as having  $n$  intercepts, one for each country (i.e., all the respondents living in the same country have the same intercept). Specifically, let  $\alpha_j = \beta_0 + \beta_2 Z_j$ . Then equation (1) becomes:

$$Y_{ij} = \alpha_j + \beta_1 X_{ij} + e_{ij} \quad (2)$$

In this case, the slope coefficient of the regression line,  $\beta_1$ , is the same for all respondents, but the intercept of the regression line varies from one country to the next. The source of variation in the intercept is the variable  $Z_j$ , which varies from one country to the next but is constant over respondents within the same country. If we assume that  $\alpha_j$ 's are all equivalent (i.e.,  $\alpha_j = \alpha_k$  for all  $j$  and  $k$ ), then equation (2) does not differ from a normal OLS (i.e., a model where all the countries can be completely “pooled” into a single population), given that we will have just one single intercept for all the observations. But this is often not the case!

Our theoretical interest is to estimate  $\beta_1$ , the effect of `life_satisf` on `demo_satisf` holding constant the unobserved country characteristics represented by  $Z$ . Note that if you do not consider explicitly  $Z$  in your model, and  $Z$  is important in affecting  $Y$  and is correlated with  $X$ , then you produce an omission bias. Moreover, given that  $Z$  is shared within each given country by all the respondents living in that country, your omission bias will inevitably create errors at the individual level within the same country that are therefore correlated among themselves (i.e. you are violating the Independence assumption!).

Of course, besides  $Z$ , you could have other variables (such as  $W$ ,  $R$ ,  $P$ , etc.) that are once again unmeasured national variables that could affect `demo_satisf`. If this happens,  $\alpha_j$  represents our ignorance about all of the systematic factors at the country level that predict  $Y$ , other than  $X$ . If these factors were known and/or measurable, they could have been included as additional covariates in the model, thus explaining the extra variation in  $Y$  and eliminating variation in  $\alpha_j$  across countries. But often they are not!

So how to deal with this problem? How to deal with such unobserved national-specific characteristics that are relatively constant within a given country? Since these variables are not directly included in the model, we can capture their effects by employing a fixed effects regression model, i.e., **using binary (dummy) variables to denote the individual countries**.

In this case, **the slope coefficient of the regression line,  $\beta_1$ , is once again the same for all respondents, but the intercept of the regression line varies from one country to the next.**

To develop the fixed effects regression model using binary variables, let  $D1_j$  be a binary variable that equals one when  $j=1$ , and equals zero otherwise, let  $D2_j$  be a binary variable that equals one when  $j=2$ , and equals zero otherwise, etc. We cannot include all  $n$  binary variables plus a common intercept, because if we do it the regressors will be perfectly multicollinear (as you already know). Hence, we have to arbitrarily omit one binary variable for one entity (i.e., one country in our case: for example,  $D1_j$ ). Accordingly, the fixed effects regression model in (2) can be written equivalently as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_2 D2_j + \gamma_3 D3_j + \dots + \gamma_n Dn_j + e_{ij} \quad (3)$$

where  $\beta_0$ ,  $\beta_1$ ,  $\gamma_2$ ,  $\gamma_3$ , ...,  $\gamma_n$  are unknown coefficients to be estimated. Equations (2) and (3) are equivalent. Equation (2) is written in terms of  $n$  country-specific intercepts, while in equation (3) the fixed effects regression model has a common intercept and  $n-1$  binary regressors. In both formulations, the slope coefficient on  $X$  is the same for all the respondents, irrespective of the country where they live

Let's go back to our example. To run a fixed effects regression we could type:

```
xi: reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use i.id
```

where the variable (id) identifies each country.

(note: If you type `xi:` before `reg` as above, Stata automatically creates a set of dummy indicators, one for each country. If you do not type `xi:` before `reg`, you get the same results but dummy indicators are not created)

```
reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use i.id
```

In this example, which is the substantial meaning of the intercept (`_cons`)? The following one: when all the other IVs are equal to zero, the expected value of `demo_satisf` for respondent *i* living in country 1 (the omitted one!) is equal to 1.693. And what about the substantial meaning of `_Iid_2=0.489`? The following one: when all the other IVs are equal to zero, the expected value of `demo_satisf` for respondent *i* living in country 2 is equal to 2.182 (i.e.,  $1.693+0.489$ ).

Then to test if all the fixed effects are jointly statistically different from zero, we can type:

```
testparm i.id
```

The `testparm` command performs a Wald test. Here, the `Prob>F` is  $< 0.05$ , so we reject the null hypothesis that the coefficients for all countries are jointly equal to zero. Therefore, country fixed effects are needed in this case.

Another (widely used) option to run a fixed effect model is also:

```
xtset id  
  
xtreg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use, fe
```

By default, Stata always omits the intercept related to the first entity (i.e., country in our case). If we want to omit the second entity, we can write:

```
char id[omit] 2  
  
xi: reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use i.id
```

As an alternative, we could have typed:

```
xi: reg demo_satisf life_satisf exp_eco exp_employ tv_use newspapers  
radio_use ib2.id
```

Now check the intercept: 2.182. **Why this specific value?** Think about that!

Another example: the happiness dataset ([happiness.dta](#)):

Let's explore the data.

There are 69,705 respondents grouped according to the country/year of the survey. On the whole, there are 80 countries/years.

```
tab country_anno
```

```
codebook country_anno
```

sodlife = respondent's satisfaction with her/his life (1-10 scale)

self = respondent's ideological left-right self-placement (0-10)

health\_bad = respondent's state of health (subjective), from 1=very good to 5=very poor

age = respondent's age (in years)

sex = respondent's sex: 1=male, 2=female

sodfin = respondent's satisfaction with financial situation

religion\_attendance = respondent's attendance of religious services (from 1=more than once a week to 8=never)

trust = trust in people: 1=respondent trusts people, 2=don't trust people

marriage = respondent is married (1) or not (0)

child = respondent has children (1) or not (0)

post\_mat4 = post-materialism index: 1=materialist, 2=mixed, 3=post-materialist

```
reg sodlife self health_bad age sex sodfin religion_attendance trust  
marriage child post_mat4
```

```
xi: reg sodlife self health_bad age sex sodfin religion_attendance  
trust marriage child post_mat4, i(country_anno)
```

You cannot run the previous model because `country_anno` is a string variable! You need first to transform such string variable into a numerical variable:

```
encode country_anno, gen(cy_num)
```

```
tab cy_num
```

Now you can run the model:

```
reg sodlife self health_bad age sex sodfin religion_attendance trust  
marriage child post_mat4 i.cy_num
```

```
testparm i.cy_num
```



Note that once the “fixed effects” are included, all the usual OLS assumption still holds intact!

What are the **limitations** with using a **fixed effects model**?

1) Once we have introduced a set of dummies, one for each country (minus one), we can explain with the remaining covariates just the variance within each country, discarding all the information (variance) between countries. In other words, any  $X_{ij}$  can only explain the variance within countries, but it cannot explain the variance/difference between countries. This is because the variance between countries is completely explained by the set of country fixed effects: i.e., the value of a given dummy for a country explains the average difference in the  $Y$  between that country and the other ones. For example, the  $\beta$  for sex in our previous equation explains the expected impact of sex on sodlife within a “general” country (by exploiting the within variance part), but it cannot explain if and why on average sodlife is higher in one or another country (i.e. the between variance part that is captured by the country-dummies). For instance, sex cannot explain the difference in sodlife between a person living in France and a person living in Germany.

2) Usually, researchers want to include in the specification important covariates of interest that does not vary within units (here, a covariate that does not vary within a country, such as economic growth). In this case, this unit-invariant predictor will be perfectly collinear with the set of unit dummy variables, making it impossible to estimate the unique effects of that variable.

3) Moreover, there can be an independent variable that exhibits extremely minimal variation within each unit (a so called “slow moving variable”). If the correlation between these slow moving variables and the unit fixed effects is high enough, this can destabilize the estimates of the effect of the independent variable.

For example, let’s say that we want to check the impact of the GDP growth (gdpgrowth) of a country on sodlife:

```
reg sodlife gdpgrowth_avg self health age sex sodfin  
religion_attendance trust marriage child post_mat4 i.cy_num  
  
xtset cy_num
```

```
xtreg sodlife gdpgrowth_avg self health age sex sodfin  
religion_attendance trust marriage child post_mat4, fe
```

Multicollinearity problems!!!

```
tab gdpgrowth_avg cy_num
```

Also take a look at this and guess the two reasons why it is not working:

```
xtreg sodlife gdpgrowth_avg self health age sex sodfin  
religion_attendance trust marriage child post_mat4 i.cy_num, fe
```