ACADEMIC YEAR 2017/2018
Università degli Studi di Milano
GRADUATE SCHOOL IN SOCIAL AND POLITICAL SCIENCES
APPLIED MULTIVARIATE ANALYSIS

Luigi Curini
luigi.curini@unimi.it

Do not quote without author's permission

Regression Diagnostics

Issues of Independence

We want that, once all covariates are considered, there are no further correlations (that is, dependence) between measures. The statement of this assumption that the errors associated with one observation are not correlated with the errors of any other observation covers several different situations. The main possibilities can be summarized in **two different scenarios**. 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).

- **Temporal dataset**

A first way in which the assumption of independence can be broken 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 observation between adjacent semesters will be more highly correlated than for observations more separated 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 likely be low in the next semester. And so on. This CAN produce also correlated error. 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 that 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

Corr(x; u_t) = 0 for all x and t, which precludes the use of the Durbin–Watson statistic 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: http://www.stanford.edu/~clint/bench/dwcrit.htm

Another possible way is using other specific tests.

An example: Using data from Klein (1950), we first fit an OLS regression of consumption on the government wage bill (minimum salary):

```
use http://www.stata-press.com/data/r11/klein
```

```
tsset yr
time variable: yr, 1920 to 1941
delta: 1 unit
```

```
regress consump wagegovt
```

If we assume that wagegov 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 0.997, 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 wagegov is strictly exogenous, we could instead use Durbin's alternative test or the Breusch–Godfrey to 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 wagegov 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. Other models – i.e., time-series regression models – allow to deal with the error process explicitly). 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).

However, if you run the following model…

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

…you do not have anymore problems with AR(1). Therefore remember: having a temporal dataset is just a necessary BUT NOT sufficient condition to have auto-correlation in your residuals. It all depends on the model you are estimating!!!

- **Hierarchical (or Multilevel) dataset**

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

Consider the case of collecting data from students in ten different elementary schools. It is likely that the students within each school will tend to be more like one another than students from different schools, that is, their errors are not independent, once controlled for all your independent variables. Or consider that you want to test the relationship between satisfaction with democracy in European countries and life satisfaction. It is likely that the respondents within each nation will tend to be more like one another than respondents from different nations, that is, 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 example. The dataset contains data on about 30,000 respondents that come from 31 countries.

We treat `demo_satisf` (i.e., satisfaction with democracy) as an interval-level variable even if in reality it is a ordinal variable (just to make one simple example).

It is very possible that the level of satisfaction for democracy within each country may not be independent, and this could lead to residuals that are not independent within countries.

Option a) Using the Cluster Option

We can use the **cluster** option to indicate that the observations in your dataset are clustered into countries (for example) and that the observations may be correlated within countries, but would be independent between countries.

Now, we can run regress with the **cluster** option (where id is the variable to identify each single country). We do not need to include the robust option since robust is implied with cluster. Note that the standard errors have changed substantially, much more so, than the change caused by the **robust** option by itself. Look for example at the coefficient for `radio_use`!

```
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)
```

As with the **robust** option, the estimate of the coefficients are the same as the OLS estimates, but the standard errors take into account that the observations within nations are non-independent. These standard errors are computed based on aggregate scores for the 31 nations, since these national level of satisfaction for democracy scores 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 could be quite larger than the OLS results. For example, if there were only 3 nations, the standard errors would be computed on the aggregate scores for just 3 nations. So NEVER use cluster s.e. when you have a low number of clusters (i.e., lower than 20 – and someone suggests 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 you model precisely this correlation (see below).

Option b) Fixed effect regression

A possible alternative to cluster standard errors is to employ so called “fixed effects”. Which is the story behind this choice? Fixed effect regression is a method for controlling for omitted variables in dataset when the omitted variables vary across entities (i.e., European countries in our previous example) but do not change over observations within a given entity (i.e., over Italians, French, Germans, etc.). 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 observation 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` (satisfaction with life variable) (X_{ij}):

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

Where Z_j is an unobserved variable that varies from one country to the next 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 but is constant over respondents within the same country, the regression model in (1) can be interpreted as having n intercepts, one for each of the respondents living in the same country. Specifically, let $\alpha_j = \beta_0 + \beta_2 Z_j$. Then equations (1) becomes:

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

In this case, the slope coefficient of the regression line, β_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 α_j 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 β_1, the effect of `life_satisf` on `demo_satisf` holding constant the unobserved country characteristics Z. Note that if you do not consider explicitly Z in your model, and Z is important in affecting Y and it 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, α_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 be included as additional covariates in the model, thus explaining the extra variation in y and eliminating variation in α_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 country instead.

In this case, the slope coefficient of the regression line, β_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 effect 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, for if we do the regressors will be perfectly multicollinear (you already know that!), so 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:

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

where $\beta_0, \beta_1, \gamma_2, \gamma_3, \dots, \gamma_n$ are unknown coefficients to be estimated. Equations (2) and (3) are equivalent: equation (2) is written in terms of n country-specific intercepts; 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 in which country they are living. How to relax this last assumption? See later our discussion about non-linear models!

Let’s go back to our example. To run a fixed effect 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. In this example, which is the substantial meaning of the intercept? The following one: the expected value of `demo_satisf` for respondent i living in country 1 (the omitted one!) when all the other IVs are = 0 is equal to 1.69. And what about the substantial meaning of _Iid_2 = .489? The following one: the expected value of `demo_satisf` for respondent i living in country 2 when all the other IVs are = 0 is equal to 2.18 (i.e., 1.69+.489).

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

```
testparm _Iid*
```

Alternatively we could have written:

```
areg  demo_satisf   life_satisf   exp_eco   exp_employ    tv_use
newspapers radio_use, abs(id)
```

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
```

Now check the intercept: 2.18. **Why this specific value?** Think about that!

Another example: the happiness dataset

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

xi: reg sodlife self health  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 in a numerical variable:

```
encode country_anno, gen (code)
tab code
```

Now you can run the model:

```
xi: reg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4 i.code

testparm _Icode*

areg sodlife self health  age sex  sodfin  religion_attendance trust
marriage  child post_mat4, absorb(code)
```

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

The **limitations** with using a fixed effect model:

1) By introducing 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 (given that such variance 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 β 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 given country than in another one (the between variance part that is captured by the country-dummies: i.e., `sex` cannot explain the difference in `sodlife` between a person living in France and a person living in Germany)

2) This has a further consequence. It is very common for a researcher to want to include in the specification an important covariate of interest that does not vary within units. In this case, the 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. Alternatively, the independent variable may exhibit extremely minimal variation within each unit (so called slow moving variables). If the correlation between these slow moving variables and the unit fixed effects is high enough, this can destabilize estimates of the effect of the independent variable.

For example, let's say that we want to check the impact of the `gdpgrowth` of a country on `sodlife`:

```
areg sodlife gdpgrowth_avg self health  age sex  sodfin
religion_attendance trust marriage  child post_mat4, absorb(code)

xi: reg sodlife gdpgrowth_avg self health  age sex  sodfin
religion_attendance trust marriage  child post_mat4 i.code
```

Multicollinearity problems!!!