ACADEMIC YEAR 2016/2017
Università degli Studi di Milano
GRADUATE SCHOOL IN SOCIAL AND POLITICAL SCIENCES
APPLIED MULTIVARIATE ANALYSIS
Luigi Curini
luigi.curini@unimi.it

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).

- **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}):

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

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:

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

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 $\beta_0, \beta_1, \gamma_2, \gamma_3, \ldots, \gamma_n$ 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!!!

Option c) Random effect regression

The random-effect model solution to the violation of independence across units is to partition the unexplained residual variance not explained by the included covariates (i.e., the error!) into two: higher-level variance between higher-level entities (i.e., countries in our example) and lower-level variance within these entities. That is, let's assume that the error term has an unobserved higher-level-specific component that does not vary within a higher-level and an idiosyncratic component that is unique to each lower-level observation. This can be achieved by having a residual term at each level: the higher-level residual is the so-called random effect. As such, a simple standard Random-effect model would be:

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

where

$$\alpha_j = \beta_0 + u_j$$

These are the micro and macro parts of the model, respectively, and they are estimated together in a combined model that is formed by substituting the latter into the former:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + (e_{ij} + u_j)$$

Where Y_{ij} is the dependent variable. In the "fixed" part of the model, β_0 is the intercept term, X_{ij} is a series of covariate(s) that are measured at the lower level with coefficient β_1. The "random" part of the model (in brackets) consists of u_j the higher-level residual for higher-level entity j, allowing for differential intercepts for higher-level entities, and e_{ij}, the respondent-level residual within country j. The u_j term is in effect a measure of "similarity" that allows for dependence, as it applies to all observations of a higher-level entity. In the random-effect model the u_j are assumed to follow a probability distribution, with parameters estimated from the data. This distribution is typically normal, with a mean of zero and a variance which describes by how much the other u_j vary around that mean.

Intuitively, the random effects model is like having an OLS model where the intercept varies randomly across countries j. Like simple OLS, the random effects model assumes that there is zero correlation between u_j and X_{ij}. If u_j and X_{ij} are correlated, the random-effects estimates are biased (see later on this point).

The nice thing about a random-effect model is that by not including a set of fixed-effects (i.e., a set of dummies one for each country that by construction explains the entire variance between countries: all higher level variance, and with it any between effects, are controlled out using the higher-level entities themselves, included in the model as a set of dummy variables), we can include a set of covariates measured at the higher level without incurring a problem of multi-collinearity (u_j is just the residual at the higher level! Not a dummy variable!). For example, we can include variable Z_j. In this case we will have:

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

where

$$\alpha_j = \beta_0 + \beta_2 Z_j + u_j$$

Therefore:

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

How to run a random model in Stata: via a Generalized least squares estimation using the xtreg command. The i() term tells STATA the variable that identifies each unique higher-level unit.

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

or via a maximum likelihood estimation:

```
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4, i(code) mle
```

You do not find any R2 here, simply because this model does not try to minimize the sum of the squared errors like the OLS does (see the Addendum).

Addendum:

GLS: In statistics, generalized least squares (GLS) is a technique for estimating the unknown parameters in a linear regression model. The GLS is applied when the variances of the observations are unequal (heteroscedasticity), or when there is a certain degree of correlation between the observations. In these cases ordinary least squares can be statistically inefficient, or even give misleading inferences.

MLE: The likelihood function is the joint probability distribution of the data, treated as a function of the unknown coefficients. The Maximum likelihood estimation (MLE) of the unknown coefficients consists of the values of the coefficients that maximize the likelihood function. Because the MLE chooses the unknown coefficients to maximize the likelihood function, which is in turn the joint probability distribution, in effect the MLE chooses the values of the parameters to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are the parameter values "most likely" to have produced the data. As with all maximization or minimization problems, this is done by trial and error. Because the MLE is normally distributed in large samples, statistical inference about the coefficients that it estimates proceeds in the same way as inference about the linear regression function coefficients based on the OLS estimator.

Given that the likelihoods can vary between 0 and 1, the logs of likelihoods (what Stata reports you) can vary between large negative numbers and 0 (the log of 1).

In the example above, when it begins the analysis, MLE finds out how well it can predict the observed values of the DV without using the IV as a predictive tool. So MLE first determined how accurately it could predict sodlife by not knowing anything else. The log-likelihood in the final iteration of the fitting constant-only model is -145194.82 and it summarizes the initial, know-nothing prediction. MLE then brings the IV into its calculations, running the analysis again – and again and again – in order to find the best possible predictive fit between years of education and the likelihood of voting. According to the logit output, MLE ran through 3 iteration, finally deciding that it had maximized its ability to predict sodlife by using the IVs as a predictive instrument. The log likelihood in Iteration 3 line is -134925.75, and it summarizes this final-step prediction.

The rho reported in the table is the so-called **intra-class correlation**, that is, a measure that tells us how much of the total variation in your dependent variable is due to a difference/ can be explained solely by differences between entities/groups (countries in our case). Formally:
$\sigma = Var(u_j)/Var(u_j + e_{ij})$, where $Var(e_{ij})$ is the variance component of happiness at the individual level and $Var(u_j)$ is the variance component at the country level. How to find such rho in the first model? In the following way:

```
di .29986227^2/(.29986227^2+1.6729226^2)
```

Sometimes it is useful to start the analysis with an a-theoretical model (so-called "Null Model") that does not include level-1 or level-2 predictors, and thus allows us to decompose the total variance in our dependent variable between the individual and country levels.

```
xtreg sodlife, i(code) mle
```

In our case we can see that approximately 13 percent of the difference in life satisfaction can be explained simply by the fact that respondents come from different countries. Moreover, the variance at level-2 is significant while the likelihood ratio test that controls if the variance at level 2 (i.e., country level) is equal to 0 can be safely rejected at standard significance levels. The null hypothesis tested by a likelihood ratio test is equivalent to the hypothesis that there is no random intercept in the model. According to our results, we can reject such null hypothesis, implying that we cannot use a pooled model and instead need a random-effects (or a fixed effects…) model to obtain reliable statistical estimates.

Alternatively (if you estimate a GLS model): recall that the random-effects model is like having an OLS model where the constant term varies randomly across higher-level units *j*. Therefore, we need to test whether there is significant variation in u_j across higher-level units. We can perform the Breusch-Pagan test by typing xttest0 after xtreg, re

```
xtreg sodlife, i(code)
xttest0
```

Our null hypothesis is that $\sigma_u^2 = 0$. Here we can reject such null hypothesis. Therefore, we can conclude that the random-effects model is preferable to the OLS model.

As already stressed, in a random model you can also include in the analysis variables at the country level:

```
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4 log_gdp unemployment_avg, i(code)
```

The **limitations** of using a random effect model:

1) Can be your higher level units considered as a random sample from a larger universe? Much easier with schools than with countries…

2) The most serious drawback of the random-effects approach is that it requires no correlation between the covariate of interest and the random effects. To illustrate why that, consider the following example: we want to investigate the effect of smoking on birth outcomes (specifically, birthweight). Smoking is a dummy variable for mother smoking during each single pregnancy (so it can also changes over the same mother). We have a 2-level structure of data: infant(s) born from (nested within) a mother. Therefore we estimate a random model. In this

framework, the coefficient that we get from the random model with respect to smoking explains both the within and the between-mothers variance. That is, our estimated coefficient for smoking represents either a comparison between children of *different* mothers (i.e., between variance at the mother level), one of whom smoked during the pregnancy and one of whom did not (holding all the other covariates constant), or a comparison between children of the *same* mother (i.e., within variance at the mother level) where the mother smoked during one pregnancy and not during the other (holding all the other covariates constant), or a mix of the two effects. According to the so called "**exogenity assumption**", smoking must be uncorrelated with u_j that is the random intercept for mother, which represents the effect of omitted mother-specific covariates on birthweight. However, mothers who smoke during their pregnancy may also adopt other behaviors such as drinking and poor nutritional intake. These variables adversely affect birthweight and have not been adequately controlled for, so that the impact of smoking on the "between variance" is likely to be an overestimate of the true effect. In contrast, each mother serves as her own control for the "within variance" so all mother-specific explanatory variables have been held constant! In this situation, by omitting cluster-level covariates we could create a situation where between-cluster relationships can differ substantially from within-cluster relationships. So that, for example, we have a significant and large negative effect of smoking on birthweight in the random model that is entirely "driven" by the between-mothers effect, while possibly being much lower (at the extreme: non-significant) in explaining the within-variance aspect (ecological fallacy!!!)

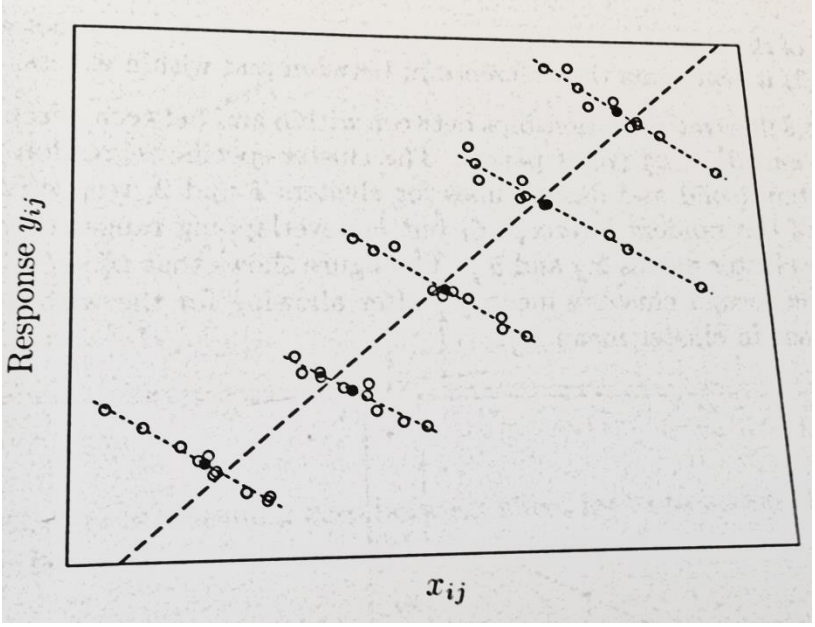

Illustration of within-cluster and between-cluster effects when the exogeneity assumption is violated

```
xtreg birwt smoke, i(momid)
xtreg birwt smoke, i(momid) fe
xtreg birwt smoke, i(momid) be
```

Addendum: What is ***xtreg, fe***? And what does?

In small datasets, it is easy to create dummy variables for each country (or each higher-level unit). In large datasets, we may have thousands of higher-level units. The number of variables in STATA is restricted due to memory limits. Also it is not very inconvenient to have results for thousands of dummy variables (just imagine how long your log file would be!).

Fortunately, STATA has a command that allows us to avoid creating dummy variables for each higher-level unit: xt is a prefix that tells STATA we want to estimate a hierarchical/multilevel data model, while the ***fe*** option tells STATA we want to estimate a fixed effects model. In OLS this is equivalent to including dummy variables to control for person(lower-level)-specific effects.

More in details, instead of including a set of dummy variables, Stata controls for idiosyncratic higher-level effects by transforming the Y and X variables:

$$Y_{ij} = \beta X_{ij} + (e_{ij} + u_j) \quad (1)$$

Taking averages of eq. (1) over higher-level units gives:

$$\overline{Y_{.j}} = \beta \bar{X}_{.j} + (\bar{e}_{.j} + u_j) \quad (2)$$

Subtracting eq. (2) from eq. (1) gives:

$$Y_{ij} - \overline{Y_{.j}} = \beta\left(X_{ij} - \bar{X}_{.j}\right) + (e_{ij} - \bar{e}_{.j}) \quad (3)$$

The key thing to note here is that the individual-specific effects (u_j) have been "differenced out" so they will not bias our estimate of β.

Compare the results of the following two equations with respects to the coefficients of `self` and `health`:

```
xtreg sodlife self health , i(code) fe
reg sodlife self health i.code
```

Now try to run the following command:

```
xtreg birwt smoke, i(momid) fe
xi: reg birwt smoke i.momid
```

You get an error! Too many "momid" (i.e., higher-level units) dummies!

What is ***xtreg, be***? And what does?

$$\overline{Y_{.j}} = \beta \bar{X}_{.j} + (\bar{e}_{.j} + u_j)$$

The averages model is sometimes called the "between" estimator because the comparison is cross-sectional between higher-level units rather than over lower-level units.

Like OLS, the between estimator provides unbiased estimates of β only if the unobservable higher-level specific component (u_j) is uncorrelated with X_{ij}.

The between estimator is also less efficient than simple OLS because it throws away all the variation over lower-level units in the dependent and independent variables.

In fact the between estimator is equivalent to estimating an OLS model on the averages for just one higher-level unit.

More generally: suppose that there is a variable z that predicts y but is not included as a covariate in the random-effects model. As a result of omitting z from the model specification, the higher or lower levels of y in unit j due to z are instead accounted for by the unit effects α_j. For there to be no bias in estimates of the coefficient on x, there must be no correlation between x and z, and hence, no correlation between x and α_j, implying no confounding due to the omitted z. On the contrary, any correlation between x and α_j can imply an omitted variable z that produces bias in estimates of β.The greater the magnitude of the correlation between x and α_j, the greater the bias in estimated of β.

Addendum: In statistics, the *bias* (or bias function) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased.

How to check for that, i.e., the existence of any correlation between u_j and X_{ij}? *Hausman test* (it only works with GLS estimation!): this takes the form of comparing the parameter estimates of the fixed effect and random model via a Wald test of the difference between the vector of the coefficient estimates of each. A significant test result is taken as evidence of a correlation between x and α_j, implying that the random-effects model should be rejected in favor of the fixed-effects model.

If u_j and X_{ij} are correlated, we should use the fixed-effects model rather than OLS or the random-effects model (otherwise the coefficients are biased).

If they are not correlated, it is better to use the random-effects model (because it is more efficient).

```
xtreg sodlife self health , i(code) fe  # see the actual
correlation between X and α_j/u_j
estimates store fixed
xtreg sodlife self health , i(code) re # see that the correlation
between X and α_j/u_j is assumed to be 0
hausman fixed ., sigmamore
```

There is a strong evidence for model misspecification since the Hausman test statistics is 37.73 with a df =2 (given that we are using in our model just two IVs). A significant Hausman test if often taken to mean that the random-intercept model should be abandoned in favor of a fixed-effects model that only utilizes within information.

Note this result is based on the specific model specification that we have tested. Therefore random effects might be appropriate for some alternate model of life satisfaction. For example, in the previous model it is pretty reasonable to suspect that u_j is correlated with `health` as long as the expectation of personal health are correlated with the quality of the national health system that chances across countries. Such quality however is not included in the model, therefore it is entirely captured by u_j , that, as a result, will be correlated with both `health` and `sodlife`. From here the omission bias!!!

Addendum: how to estimate the *goodness of fit* of two or more different models? With a random model you do not have anymore your beloved R2. Is it a problem? Not at all! Remember what we already told about the substantial meaning of R2. So what do to? Well, you compare the likelihood of the models through a likelihood-ratio test!

The likelihood-ratio (LR) test is one of the three classical testing procedures used to compare the fit of two models (the other two being AIC and BIC), one of which, the constrained model, is nested within the full (unconstrained) model. Under the null hypothesis, the constrained model fits the data as well as the full model. The LR test requires one to determine the maximal value of the log-likelihood function for both the constrained and the full models. More in details, you have to estimate the following: -2*(log likelihood of the constrained model- log likelihood of the full model) and then you have to compare the result with the ChiSquared distribution to understand if such difference is significant or not (see the file: "Critical values for the chi squared distribution.pdf"). Otherwise you can rely to the Stata command lrtest (see below).

Remember, therefore, that such test work only when the statistical model has a likelihood (also an OLS has a likelihood! But a GLS no!). And remember that such test works only for nested models!!!! Usually, this means that both models are fit with the same estimation command (for example, both are fit by xtreg, with the same dependent variables) and that the set of covariates of one model is a subset of the covariates of the other model.

with xtreg MLE

```
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4, i(code) mle
estimates store full
estat ic
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child , i(code) mle
estat ic
```

```
lrtest full .
```

Alternatively:
Likelihood ratio and then look at the chi-squared table! With 1 d.f. given that the full model has 1 IV more than the nested one
di -2*(-134927.45 - -134925.75)

```
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4, i(code) mle
estimates store full
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child , i(code) mle
estimates store half_full
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  , i(code) mle
lrtest full  .
lrtest half_full .
lrtest full half_full
```

with reg

```
reg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4
estimates store full
estat ic
reg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child
lrtest full .
estat ic
```

As an alternative you can employ information criteria such as AIC (Akaike's information criteria) or BIC (Bayesian information criteria).

Information criteria: a statistic used to estimate which models is the best one. The AIC (BIC) is not a test on the model in the sense of hypothesis testing, rather it is a tool for model selection. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. From the AIC value one may infer that e.g the top three models are in a tie and the rest are far worse, but one should not assign a value above which a given model is "rejected".

Unlike likelihood-ratio, the models need not be nested to compare the information criteria. Because they are based on the log-likelihood function, information criteria are available only after commands that report the log likelihood.

Akaike's) information criterion is defined as
AIC = -2 lnL + 2k
where lnL is the maximized log-likelihood of the model and k is the number of parameters estimated.
Some authors define the AIC as the expression above divided by the sample size.

Bayesian information criterion is another measure of fit defined as
BIC = -2 lnL + k lnN
where N is the sample size.

Addendum: The *chi squared distribution* is the distribution of the sum of m squared independent standard normal random variables. This distribution depends on m, which is called the degrees of freedom of the chi-squared distribution. For example, let Z1, Z2 and Z3 be independent standard normal random variables. Then $Z1^2+Z2^2+Z3^2$ has a chi-squared distribution with 3 degrees of freedom.

Summing up:

So what should we use? Random or fixed effects? It depends, as always, on your theoretical aims…Is your theoretical model (mainly) dealing with what happens within a general country? Are you (mainly) interested in that? If yes, then it's ok using a fixed effect model (i.e., explaining the variance within a country and discarding the variance between countries, that is entirely captured by the fixed effects).

Any time one's theoretical model does not dictate a particular specification, we can move to empirical evaluation. That is, we can investigate empirically which model offers better inferences about the quantities of interest. By using an Hausman test, as already discussed, or by considering efficiency loss. How much is the within country variation compared to the between country variation? Run an xtreg, re model and check for that! If the within country variation in Y is, for example, four times the between-countries variation, then you face low efficiency loss in using fixed effect (and discarding between country-variance). Moreover remember that every time there are covariates having the same within- and between-effects, we obtain more precise estimates of these coefficients by exploiting both within- and between-cluster information.

If you have enough cluster, cluster s.e. will produce results quite similar to a random model.

```
xtreg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4 log_gdp unemployment_avg, i(code)
mle

reg sodlife self health  age sex  sodfin  religion_attendance
trust marriage  child post_mat4 log_gdp unemployment_avg,
cluster(code)
```

But remember that with cluster s.e. you do not model anything with respect to the issue of independence. You just try to cure it…

The magical world of the fixed-effects vs. random-effects models, including multilevel models, is incredibly rich! This was just an introduction. See for example:

Tom S. Clark and Drew A. Linzer, Should I Use Fixed or Random effects?, *Political Science Research and Methods*

Andrew Bell and Kelvyn Jones, Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data, *Political Science Research and Methods*

Steenbergen, M.R., and B.S. Jones, Modeling Multilivel Data Structures, *American Journal of Political Science*

Rabe-Hesketh, S., and S. Anders, *Multilevel and Longitudinal Modeling Using Stata*

- **Temporal dataset**

A second 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!!!