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Do not quote without author's permission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.

- **Hierarchical (or Multilevel) dataset**

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!!!)

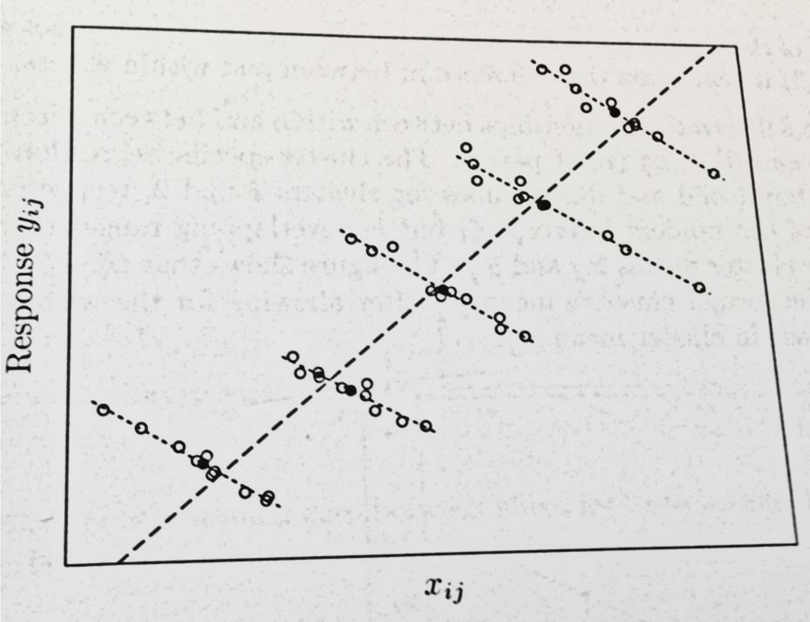

Illustration of within-cluster and between-cluster effects when the exogeneity assumption is violated

```
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:

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

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

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

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

$Y_{ij} - \overline{Y_{.j}} = \beta\left(X_{ij} - \bar{X}_{.j}\right) + (e_{ij} - \bar{e}_{.j})$ (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 αj/uj
estimates store fixed
xtreg sodlife self health , i(code) re # see that the correlation
between X and αj/uj 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*