

## Regression Diagnostics

### Issues of Independence

Independence means that the errors associated with one observation are not correlated with the errors of any other observation. We want that, once all covariates are considered, there are no further correlations (that is, dependence) between measures. As we already know, dependence may arise when our dataset is hierarchical (or multilevel). 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

The random-effects model solution to the violation of independence across units is to partition into two parts the unexplained residual variance. In other words, the random-effects model partitions the variance that is not explained by the included covariates (i.e., the error!) into the following two parts: 1) higher-level variance between higher-level entities (for example, countries), and 2) lower-level variance within these entities (for example, the citizens living in the same country). In other words, we assume that the error term has an unobserved higher-level-specific component that does not vary within a higher-level entity, 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, for individual  $i$  and high-level entity  $j$ , a simple standard random-effects model would be:

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

where:

$$\alpha_j = \beta_0 + u_j$$

These two formulas represent the micro and macro parts of the model, respectively. 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. The **“fixed” part** of the model includes  $\beta_0$ , which is the intercept term, and  $X_{ij}$ , which represents a series of covariate(s) that are measured at the lower level with coefficient  $\beta_1$ . The **“random” part** of the model consists of the two terms in brackets:  $u_j$ , which is the higher-level residual for higher-level entity  $j$  (this allows for differential intercepts for higher-level entities), and  $e_{ij}$ , which is 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 in a higher-level entity. In the random-effects 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 that 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-effects model is that, since we do not include a set of fixed effects – i.e., a set of dummies, one for each country, that by construction explain the entire variance between countries (all higher-level variance, and with it any between-countries 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 multicollinearity. This is because  $u_j$  is just the residual at the higher level, and not a dummy variable. For example, we can include variable  $Z_j$ , which measures some characteristic of countries (e.g. GDP growth, the type of political regime, the quality of democracy, etc.). 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:

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

Here’s a nice way to understand and visualize random-effects models (especially more sophisticated ones):

<http://mfviz.com/hierarchical-models/>

How to run a random model in STATA? By using the `xtreg` command, where the `i()` term tells STATA which is the variable that identifies each unique higher-level unit. When using the `xtreg` command, we can choose which technique to use in order to estimate random effects: a Generalized Least Squares (GLS) estimation or a Maximum Likelihood Estimation (MLE). Let's open the happiness dataset (`happiness.dta`). If we do not specify the `mle` option, we use the GLS estimation:

```
xtreg sodlife self health_bad age sex sodfin religion_attendance  
trust marriage child post_mat4, i(cy_num)  
  
xtreg sodlife self health_bad age sex sodfin religion_attendance  
trust marriage child post_mat4, re  
  
xtreg sodlife self health_bad age sex sodfin religion_attendance  
trust marriage child post_mat4
```

If we specify the `mle` option, we request the maximum likelihood estimation:

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

You do not find any  $R^2$  here, simply because this model does not try to minimize the sum of the squared errors like the OLS does.

#### **Addendum: GLS and MLE**

**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 (heteroskedasticity), 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 in such a way as to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are those parameter values which are “most likely” to have produced the data. As with all maximization or minimization problems, this is done by a trial and error process. 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 (that STATA reports to 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 IVs as a predictive tool. So, MLE first determined how accurately it could predict `sodlife` without knowing anything else. The log-likelihood in the final iteration of

the fitting constant-only model is -145194.82 and summarizes the initial, know-nothing prediction. MLE then brings the IVs into its calculations, running the analysis again – and again and again – in order to find the best possible predictive fit between `sodlife` and the independent variables. 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 summarizes this final-step prediction.

When running `xtreg` with the `mle` option, the **rho** that is reported in the regression table is the so-called **intra-class correlation**. Intra-class correlation is a measure that tells us how much of the total variation in your dependent variable is due to a difference – i.e. can be explained solely by differences – between entities/groups (countries in our case). Formally:

$$\rho = \text{Var}(u_j) / \text{Var}(u_j + e_{ij})$$

where  $\text{Var}(e_{ij})$  is the variance component of life satisfaction at the individual level, while  $\text{Var}(u_j)$  is the variance component at the country level. In this case, the value of rho is .0563366, implying that 5.63 percent of the total variation in life satisfaction is explained by differences between countries. The rest of variation in life satisfaction (94.37 percent) is explained by differences between individuals.

The rho is also provided in the model based on GLS, being computed as follows:

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

It is often useful to start the analysis with an a-theoretical model (so-called “Null Model”) that does not include level-1 or level-2 predictors, including instead only the random intercepts. This allows us to decompose the total variance in our dependent variable between the individual and country levels.

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

In this case, we can note that approximately 13 percent of the variation in life satisfaction can be explained simply by the fact that respondents come from different countries. Moreover, the variance at level 2 (i.e., country level) is statistically significant (look at the 95% confidence intervals). In addition, the likelihood-ratio test controls if the hypothesis that 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, here we can reject such null hypothesis, implying that we have to take into account the hierarchical structure of the data. Hence, we cannot use a pooled model, but we 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`:

```
xtreg sodlife, i(cy_num)
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 discussed, in a random-effects model you can also include in the analysis variables at the country level:

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

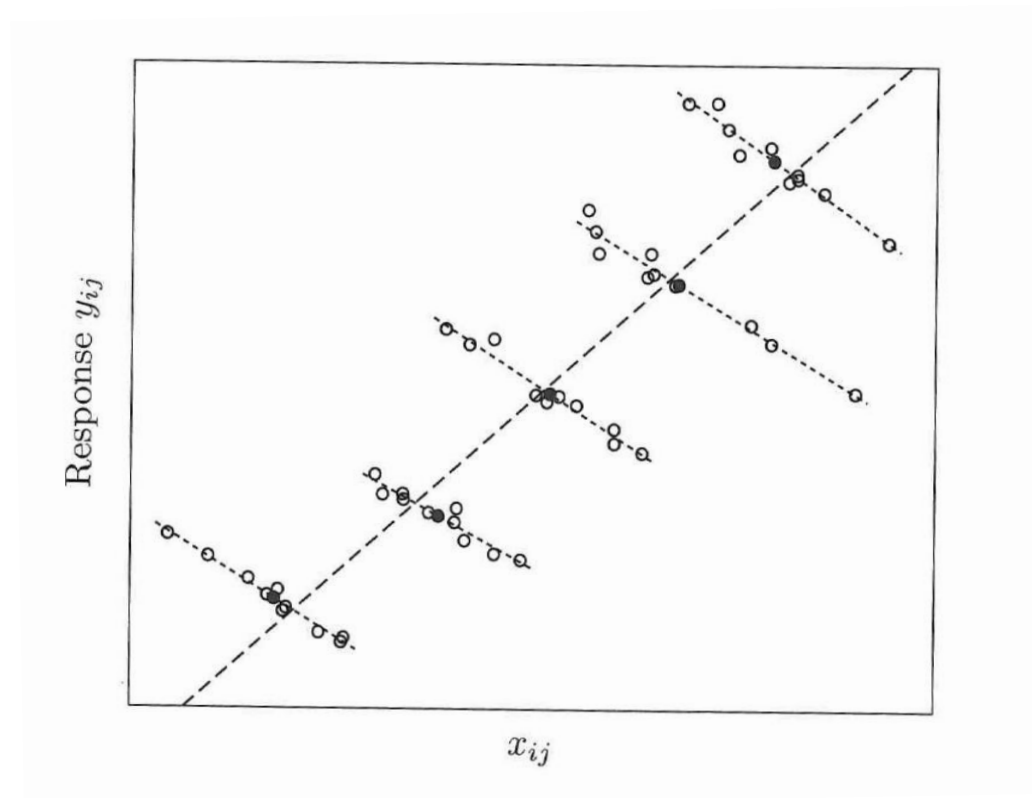
What are the **limitations** of using a random-effects model?

1) Can be your higher-level units considered as a random sample from a larger universe? Often this is not the case. For example, it is 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 covariates of interest and the random effects. To illustrate why that, open the smoking.dta dataset and consider the following example: we want to investigate the effect of smoking on birth outcomes – specifically, infant’s weight (birwt). Variable smoke is a dummy variable for mothers smoking during each single pregnancy (so it can also change 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-effects model. In this framework, the coefficient that we get from the random model with respect to smoking (i.e. the coefficient on smoke) explains both the within-mother and the between-mothers variance. In other words, the estimated coefficient on smoke 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 “**exogeneity assumption**”, smoke must be uncorrelated with  $u_j$  which is the random intercept for the high-level group (mother) and represents the effect of omitted high-level group-specific (mother-specific) covariates on the DV (infant’s weight). However, mothers who smoke during their pregnancy may also have adopted other behaviors such as drinking and poor nutritional intake. These variables adversely affect infant’s weight 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 (i.e. covariates measured at the level of groups – here, mothers) 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 infant’s weight 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!!!)

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

The following graph illustrates within-cluster and between-cluster effects when the exogeneity assumption is violated.



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  $\alpha_j$ . 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  $\alpha_j$ , implying no confounding due to the omitted  $Z$ . On the contrary, any correlation between  $X$  and  $\alpha_j$  can imply an omitted variable  $Z$  that produces bias in estimates of  $\beta$ . The greater the magnitude of the correlation between  $X$  and  $\alpha_j$ , the greater the bias in estimates of  $\beta$ .

How to check for the correlation between  $u_j$  and  $X_{ij}$ ?

Use the **Hausman test** (note it only works with GLS estimation!). The Hausman test compares 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 existing between  $X$  and  $\alpha_j$ , 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).

To see the actual correlation between  $X$  and  $\alpha_j/u_j$ :

```
xtreg sodlife self health_bad, i(cy_num) fe
estimates store fixed
```

To see that the correlation between  $X$  and  $\alpha_j/u_j$  is assumed to be 0:

```
xtreg sodlife self health_bad, i(cy_num) re
estimates store random
hausman fixed random, sigmamore
```

The latter command tests the appropriateness of the random-effects estimator. In this case, 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 is 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 alternative model of life satisfaction. For example, in the previous model it is pretty reasonable to suspect that  $u_j$  is correlated with  $health\_bad$  as long as the expectation of personal health are correlated with the quality of the national health that changes 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  $health\_bad$  and  $sodlife$ . From here the omission bias!!!

### Summing up:

So what should we use? Random effects 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 using a fixed effects model is ok (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 a Hausman test, as already discussed, or by considering efficiency loss. How much is the within-country variation compared to the between-countries 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 effects (and discarding between-countries 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 clusters, clustered standard errors will produce results quite similar to a random model.

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

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

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

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

- Clark, T.S. and Linzer, D.A. (2015), Should I Use Fixed or Random effects?, *Political Science Research and Methods*, 3(2): 399-408.
- Bell, A. and Jones, K. (2015), Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data, *Political Science Research and Methods*, 3(1): 133-153.
- Steenbergen, M.R. and Jones, B.S. (2002), Modeling Multilvel Data Structures, *American Journal of Political Science*, 46(1):218–37.
- Rabe-Hesketh, S. and Skrondal, A. (2012), *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.