WDI Report

Charley Ferrari, Christina Taylor, David Stern

May 10, 2016

Methodology

Panel data is defined as data observed longitudinally over time and across various groups. This sort of structure changes the purpose of a model.

More traditional OLS models, and even GLS models taking time into account, can have a more predictive purpose. You have a clearly defined response variable, and you collect as many exogenous variables across as many observations as possible to describe the variation in the response variable. These sorts of models can be used for both prediction or inference, you can use it to predict a response variable given exogenous variables, or you can get an idea of how your response variable is being affected by other variables.

Panel models are more important for inference. They are meant to describe what is happening to members of your group, and depending on the type of model, meant to infer characteristics of the larger group you're sampling from. You can use panel models to predict the futures of the members within your group, but it won't be as useful to predict what might happen to a new member.

Variable selection can also be looked at in this lense. If the goal is prediction, the goal is to end up with the most significant variables that describe the greatest percentage of variation in the response. If the goal is inference, you might be more interested in describing how a particular exogenous variable affects a certain response. You can ask similar questions, but the goal of the study is to find out how the particular variable x is affecting the response variable y.

The fact that Panel data is grouped makes these questions more interesting. The question isn't just how a particular variable x affects y, but what sort of variation this effect has among the different members of the group.

We have taken a very inferential approach to modeling this data, and thus are not interested in variable choice and efficient models. Rather, we are interested in building a framework that lets us gain insights both into the variable effect, and the variation of that effect over the group.

After selecting our variable, we are defining the following framework to analyze its effect:

- Exploratory Data Analysis - primarily visual analysis to get a preliminary idea of how the members of the group compare.
- ANOVA - Considering only the response variable across various groups, and focusing on the f-test to decide whether or not the means are significantly different from eachother.
- Intra-class correlation coefficient: The ICC is defined mathematically as $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$. It is the ratio of the variance in y due to fixed effects over the total variance in y. An ICC close to 1 implies that the variance is mostly due to variation between groups, while an ICC close to 0 implies that the variance is mostly occurring within groups.
- Naive Model: Build a naive model that totally ignores the groups and time periods.
- Fixed Effects Model: Build a preliminary fixed effects model, that takes into account your groups as categorical variables. See what effect this has on the significance and value of the coefficient for your independent variable.
- Random Effects Model: Build a preliminary random effects model, that considers the groups you're looking at as sampled from a general population. Instead of defining your categories as dummy variables, this gives you measures of the variance of the population your categories are chosen from.

- Hausman Test: One of the key assumptions of the Random Effects model is that the groups (as a categorical variable) is uncorellated with any other independent variables. This concept should be considered on its own, but a Hausman test mathematically determines if this is true.
- Durbin Watson Test: Alok Bhargava (Bhargava 2001) recommends using the Panel Durbin Watson Test. Defined similarly to the standard Durbin Watson test, the panel test statistic is:

$$d = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}(e_{it} - e_{it-1})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} e_t^2}$$

Bhargava defines this slightly differently in terms of the y's:

$$d = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it} - y_{it-1})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y_{it}})}$$

In the plm package, this can be tested using the pdwtest function.

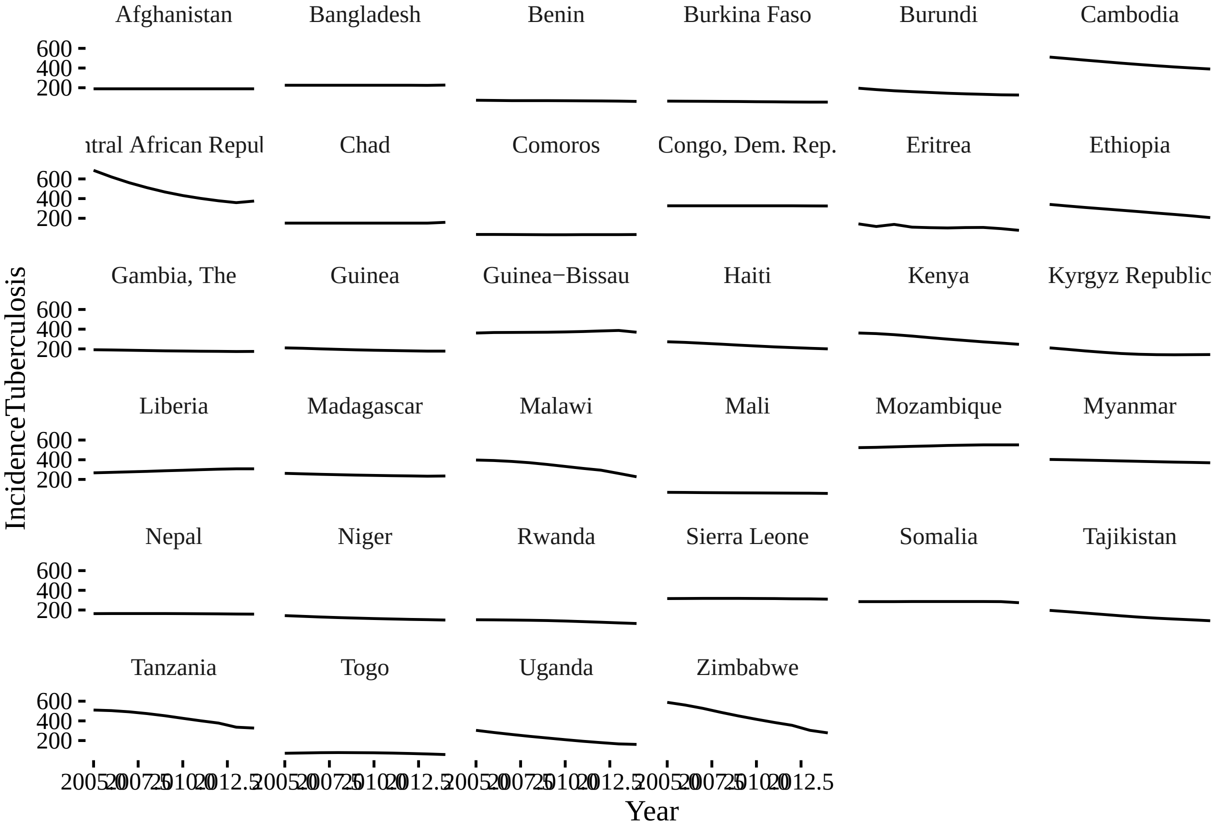
- Panel GLS Model: Based on the results of the Durbin Watson Test, we can choose whether to implement a GLS model for our data.

Example: Tuberculosis Incidence

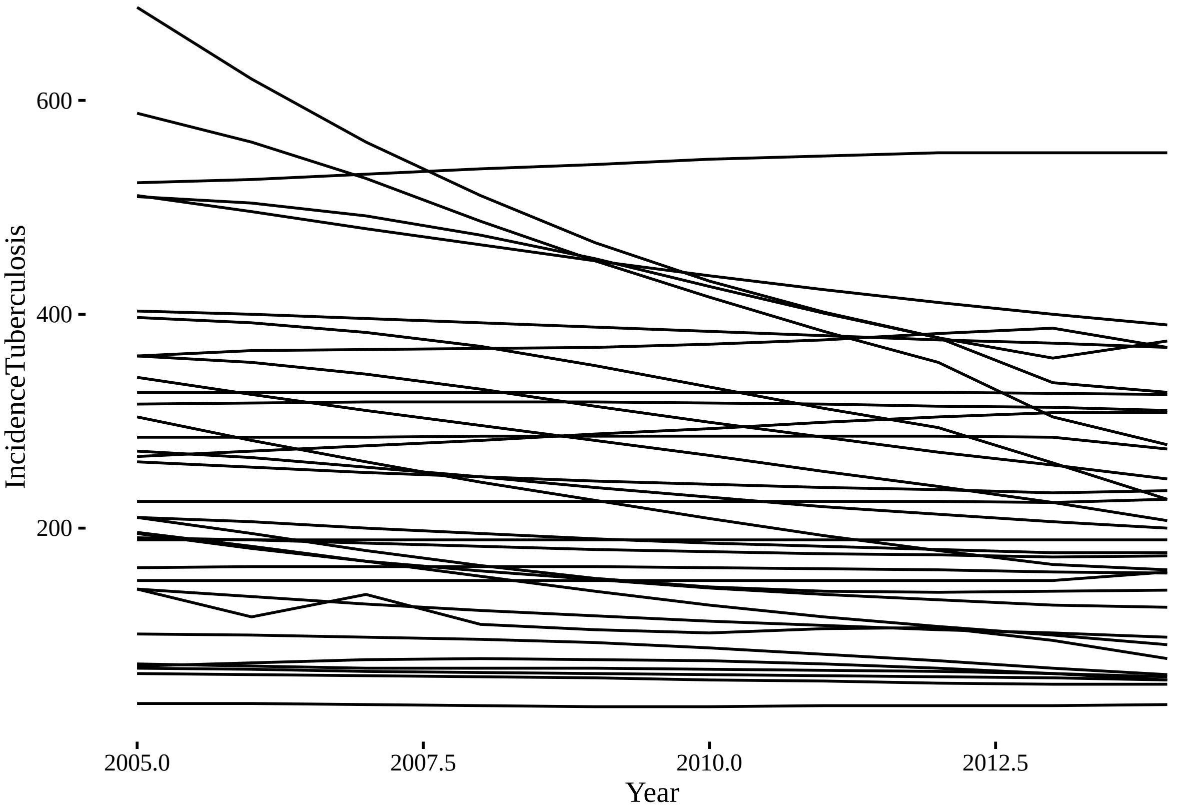
As an illustrative example, we will perform the above steps looking at the Tuberculosis Incidence rate. Our goal is to see what sort of effects the amount of aid received has on the incidence of tuberculosis over time.

First, we will have to filter our data to only look at low income countries: countries that are most likely to receive aid. We will further filter our country list to include only those that have data on aid from 2005 to 2014.

First we will perform some exploratory data analysis, looking at a faceted chart of change in Tuberculosis Incidence over time:



And the plots grouped together, to get a clearer idea of how the countries compare to eachother:



This preliminary exploratory analysis seems to suggest that there is more variation in the incidence of tuberculosis between countries than within them.

ANOVA confirms this view, giving us a p-value $< 2 \times 10^{-16}$, and confirming the alternative hypothesis that at least one of the means are different. The ICC confirms this view: 0.9323198 is closer to 1, indicating that the variance between countries accounts for the majority of the total variance.

The first model we will look at is a naive one, in the form:

$$IncidenceTuberculosis = \beta_0 + AidPerCapita \times \beta_1$$

This model gives us an extremely low r-squared of 0.00108, with a negative adjusted r-squared of -0.00187. Below is a table of the variable statistics:

Variables	Estimates	Std.Error	t.value	pr.t
Intercept	243.84	12.61	19.34	0.000
AidPerCapita	-0.10	0.17	-0.60	0.546

The estimate for AidPerCapita is not significantly different from 0, overall pointing to a very weak relationship when not including the country effects.

Next, we will build a fixed effects model. This is equivalent to adding country as a categorical variable in our model. For the 34 countries we're looking at, this would be the same as adding 33 dummy variables. The form of the fixed effects model in this case is:

$$IncidenceTuberculosis = \mu + c_i + AidPerCapita \times \beta$$

Adding fixed effects, the Estimate for AidPerCapita is now -0.23, with a p-value of 0.0026. The estimate has stayed the same sign, while our p-value has become more significant with the addition of country based fixed effects.

The random effects model assumes that the effect of Country is due to a random variable. We wouldn't estimate the variables directly like in a fixed effect model, but would end up estimating the parameters of the random variable. The random effect has to have a mean of 0, so the important parameter being estimated is the variance. Our goal in this model is to get an idea of the distribution of the country random effects.

Our β for AidPerCapita remains similar at -0.23, with a similar p-value. The variance of the idiosyncratic random effects is 1278.03, while the individual random effects is 18649.36, once again confirming our findings about the variance within versus between countries.

This model assumes however that the fixed effect isn't correlated with AidPerCapita. We can use a Hausman Test to find out if that's true. With a p-value of 0.78, the Hausman test confirms the alternative hypothesis, and leads us to conclude that the random effects model is exhibiting omitted variable bias.

Lastly, we can calculate the Durbin-Watson Panel Test statistic. At 0.744, the p-value is extremely small and we assume the alternative hypothesis that there is positive serial correlation. Using the pggls function, we can perform a "within" GLS model, indicating that we want to include the fixed effects of the countries.

This model drastically improves the R-Squared: giving us a value of 0.9375. It also takes away the significance of our AidPerCapita variable: giving us a lower estimate than we saw in previous models (-0.0094) and a high p-value of 0.58.

Taken together, these results don't give us the predictive power other OLS models may give us, but it gives a rich picture of how Aid might be affecting the Incidence of Tuberculosis.

Our exploration of naive, fixed, and random effects models indicated that experience varies greatly between countries. It suggests that more research should be done in what sort of underlying variables make these countries different if we want to come up with general theories of development.

More importantly, the Durbin Watson test indicated problems with autocollinearity, which suggested the need for a GLS. This indicates that autocollinearity is the most major factor, and accounts for the most improvement in the R-Squared.