WDI Report

Charley Ferrari, Christina Taylor, David Stern

May 10, 2016

Abstract

Based on the World Bank's World Development Indicators, we sought regression models that can predict socio-economic development. We chose three categories of development metrics: health, education, and technology; the response variables are TB incidence, primary school enrollment, and internet users per 100 people. Missing data, non-linearity, serially correlated errors, multicollinearity and differences among countries challenged our familiar OLS models and variable selection techniques. To capture both the variance between subjects as well as the changes over time, we performed panel regression analysis with both random and fixed effects, as well as general least square fit. We tested the models and found that the country factor absorbed most of the variance; the errors are serially correlated. Including country fixed effects and GLS to correct for autocorrelation lead to a vast improvement in our model's explanatory power. Our inconclusive conclusion is that there is no "one true model" that can predict world development, due to the inherent variance of countries and changes over time. We are much better off limiting our scope to specific countries and time frame, and perform regression on a case by case basis.

Key Words

Panel regression, longitudinal data, random and fixed effects, serial correlation, heteroskedasticity

Literature review:

Other research on related topics inspired our choice of predictors. Just like the authors of *Evaluating the Impact of Foreign Aid on Economic Growth: A Cross-Country Study*, we chose foreign aid as a predictor for growth. *Determinants of Enrollment in Primary Education: A Case Study of District Lahore* also helped us focusing on indicators linked to enrollment in primary education decisions. Furthermore, we also corrected for serial correlated errors and limited our time frame as well as region.

Our main difference from the above research is our panel regression methodology. Contrary to the first research, we make no assumption that a predictor's contribution to development is similar across one income group. Instead, we analyzed country fixed effects to explain the subject level variance. Unlike the second research, we built models that could apply to more than one specific location, using country as fixed effect estimator, dummy variable, or fitting different intercept/slopes.

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 through 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 - primarly 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.

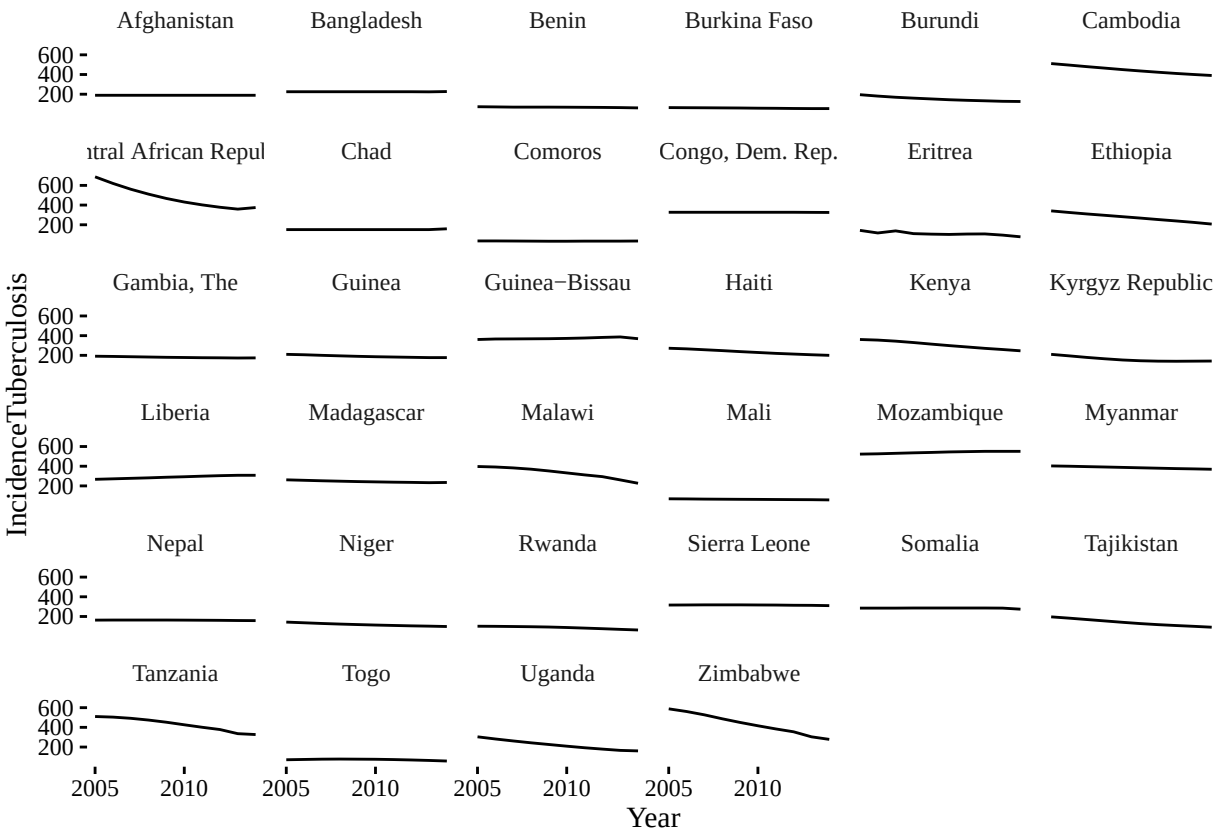
- Panel GLS Model: Based on the results of the Durbin Watson Test, we can choose whether to implement a GLS model for our data.

Example: Tuberculosis Incidence

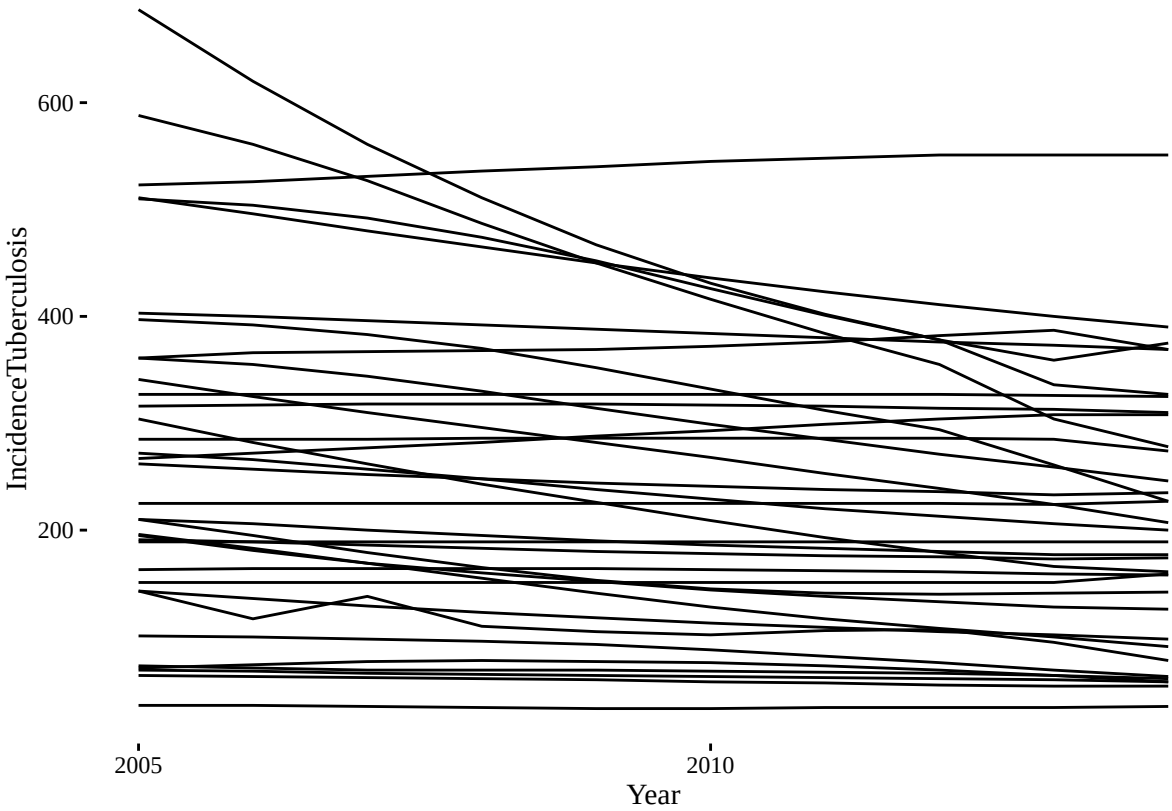
As an illustrative example, we will perform the above steps looking at the Tuberculosis Incidence rate. Our goal is to see what sort of effects the amount of aid received has on the incidence of tuberculosis over time.

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.

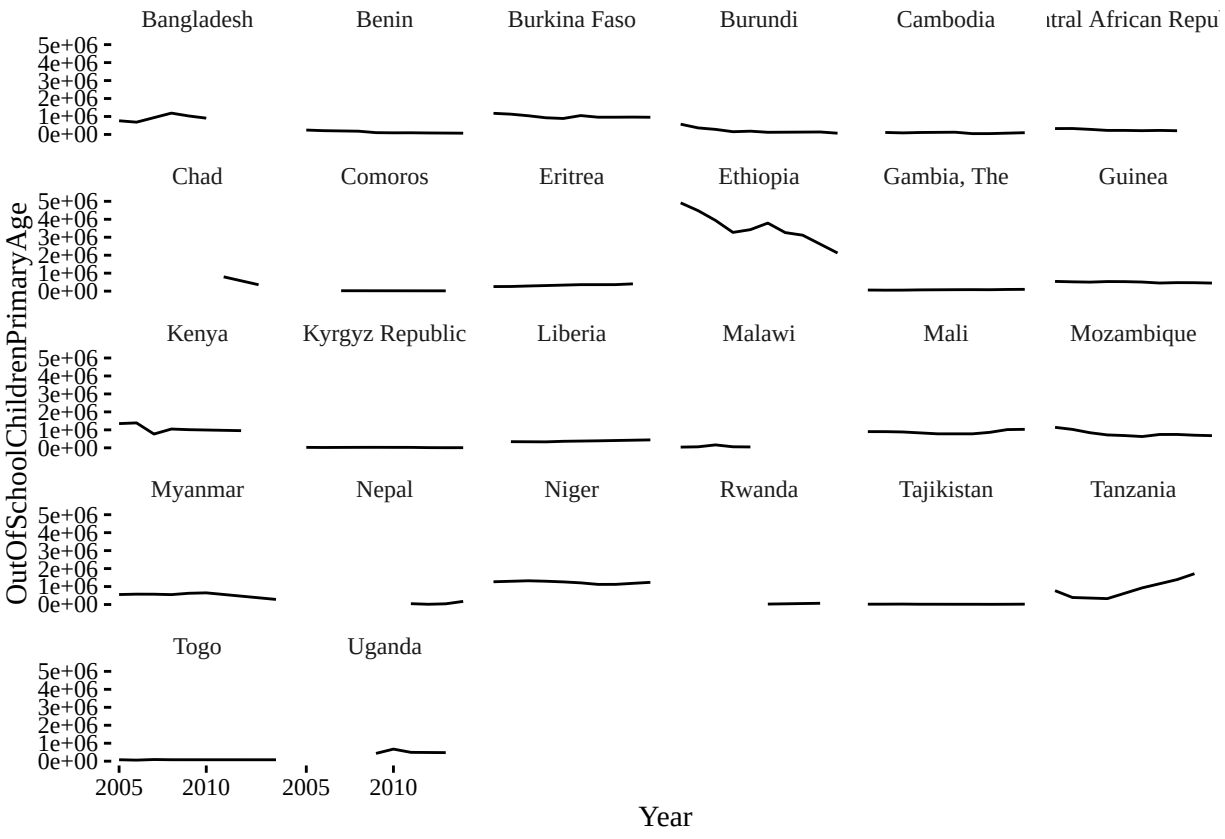
Our exploration of naive, fixed, and random effects models indicated that experience varies greatly between countries. It suggests that more research should be done in what sort of underlying variables make these countries different if we want to come up with general theories of development.

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.

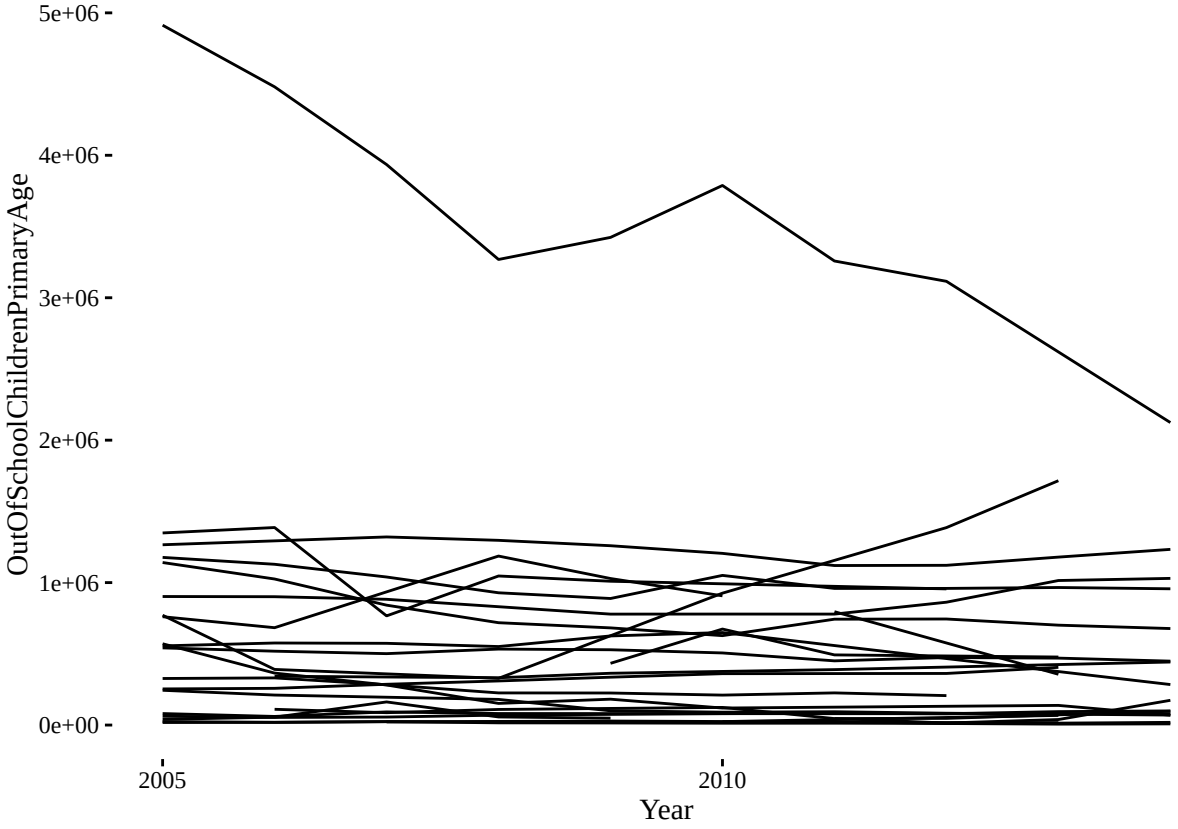
Example: Enrollment in Primary Education

We used the same methods to explore possible predictors that affect development outcomes in education. Given the high numbers of missing values and inconsistency in reporting across time and geography, we first combed through the dataset for education indicators with a low percentage of missing values. Inspired by the study *Determinants of Enrollment in Primary Education: A Case Study of District Lahore*, we considered predictors that were directly related to education (distance to school, literacy ration) along with those linked to household characteristics (ratio of dependents to working members, education expenses, family size). We found that *Out of school children of Primary Age (both sexes)* had the most data of the 25 indicators we hand-picked from more than 1300 WDI indicators, so we selected this as our response variable for education.

In our exploratory data analysis for low-income countries, it appears that there was some variation within a few countries, but that most of the variation seems to exist between countries. Guinea-Bissau, Zimbabwe, and Sierra Leone were excluded from the set of low-income countries as they only have one data point each for the period 2005-2015.



We see evidence for both sources of variation better when we group these plots the countries together:



We find more evidence that most of the variation for *OutOfSchoolChildrenPrimaryAge* can be found between countries when we group the data by country and evaluate it is as the sole predictor. The ANOVA test also indicated here, with a very significant p-value $< 2 \times 10^{-16}$, that there that there are different means between groups. The ICC value here was 0.991, demonstrating that grouping by country accounts for nearly all of the variance and that we should be leaning towards a fixed effects model.

We examined two naive models, each with *GDPperCapita* and *AidPerCapita* as the sole predictors of *OutOfSchoolChildrenPrimaryAge* for low-income countries. The coefficient for *GDPperCapita* in model was quite large, positive, and had a very significant p-value. This results is somewhat counterintuitive, as we would not expect *OutOfSchoolChildrenPrimaryAge* to increase with *GDPperCapita*. the adjusted r-squared value for this model is 0.099.

$$OutOfSchoolChildrenPrimaryAge = \beta_0 + GDPpc \times \beta_1$$

Variables	Estimates	Std.Error	t.value	pr.t
Intercept	419886.73	73863.30	5.68	1.00e-07
GDPperCapita	79714.11	17809.18	4.48	1.38e-05

The coefficient for *AidPerCapita* was more intuitive, as it appeared to be a large negative number. We would expect the number of out of school children to decrease as foreign aid increases. The p-value for the *AidPerCapita* coefficient appeared to be significant, but the r-squared value for the model as whole was 0.018. Unfortunately, it does not seem like either of these predictors explain a significant amount of variance.

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

Variables	Estimates	Std.Error	t.value	pr.t
Intercept	805508.09	110663.51	7.28	0.00
AidPerCapita	-3628.25	1754.09	-2.07	0.04

When we build fixed effects models for *AidPerCapita* and *GDPperCapita*, the coefficients change and become statistically insignificant. The coefficient for *AidPerCapita* increases from -3628.25 to -1038.7 and the coefficient for *GDPperCapita* drops from 79,714 to 5,186. Unfortunately, both of these indicators appear to be poor predictors of *OutOfSchoolChildrenPrimaryAge* in our panel estimators with fixed country effects.

We also built two simple random effects models that included the *AidPerCapita* and *GDPperCapita* indicators and performed a Hausman test with each of their respective fixed effects model. The p-values for the Hausman test were both greater than a 0.05 level of significance, so we rejected the null hypotheses that the β_1 coefficient for the random effects models were inconsistent. Since we were not able to improve the models with *AidPerCapita* and *GDPperCapita* as preditors, we looked for other indicators that explain more of the variance in *OutOfSchoolChildrenPrimaryAge*. Three indicators we found to explain some of the variance of the number of out-of-school children within countries were: the percentage of the population under the age of 14, *PctPopUnder14*; the percentage of children enrolled in pre-primary education, *PrePrimaryEnrollment*; and the fertility rate, *FertilityRate*. Fitting panel estimators with fixed effects with each of these indicators as predictors, we found that each explained a small, but considerable amount of variance.

Population Under-14

The panel model with *PctPopUnder14* as the predictor has a large, positive coefficient and is statistically significant. We can intrepret the coefficient as meaning the number of out of school children of primary age will increase as the percentage of the population increases. This makes sense for low-income countries that are experiencing population booms but cannot keep pace with providing adequate education. The model has a r-squared value of 0.14, adjusted r-squared of 0.12 and F-statistic of 24.71.

$$OutOfSchoolChildrenPrimaryAge = \beta_0 + PctPopUnder14 \times \beta_1$$

Variable	Estimate	Std.Error	t.value	pr.t
PctPopUnder14	124915	25129	4.97	1.8e-06

Pre-Primary Enrollment

The panel model with *PrePrimaryEnrollment* as the predictor has a large, negative coefficient and is also statistically significant. This coefficient also makese sense intuitively, as the number of out-of-school primary children should decrease as the number of children enrolled in pre-primary education increases. The model has a r-squared value of 0.21, adjusted r-squared of 0.17 and F-statistic of 29.414.

$$OutOfSchoolChildrenPrimaryAge = \beta_0 + PrePrimaryEnrollment \times \beta_1$$

Variable	Estimate	Std.Error	t.value	pr.t
PrePrimaryEnrollment	-28114.7	5183.9	-5.42	3e-07

Fertility Rate

The panel model with *FertilityRate* as the predictor has a large, positive coefficient and is statistically significant. The coefficient here is positive, and follows similar logic to that of *PctPopUnder14*, as low-income countries with high fertilty rates likely have trouble providing adequate education to a burgeoning population of children. This model has a r-squared value of 0.14, adjusted r-squared of 0.12 and F-statistic of 25.43.

$$OutOfSchoolChildrenPrimaryAge = \beta_0 + FertilityRate \times \beta_1$$

Variable	Estimate	Std.Error	t.value	pr.t
FertiltyRate	413265	81959	5.04	1.3e-06

Multiple Predictors

Since the three predictors each explained a small percentage of variance, we experimented adding them in different combinations in fixed-effects panel models. The most powerful model was the combination of *PctPopUnder14* and *PrePrimaryEnrollment* as predictors. The coefficients for both of the predictors had the same sign as the simple models, and were both statistically significant. This model has a r-squared value of 0.25, adjusted r-squared value of 0.20, and F-statistic of 18.30.

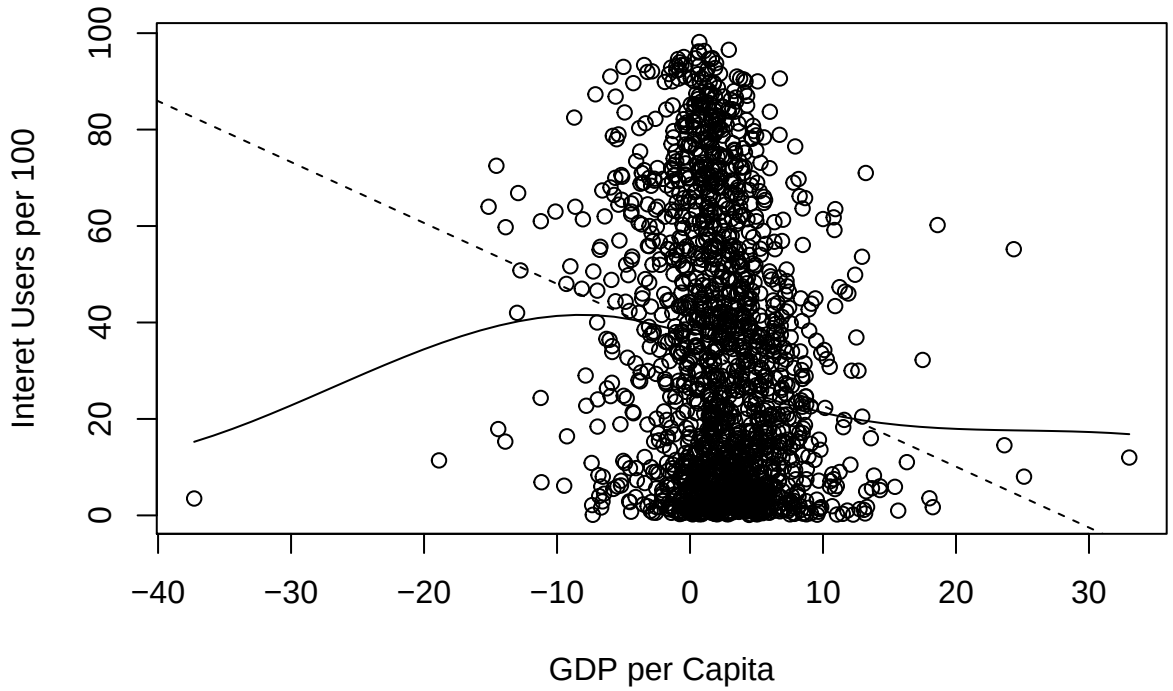
$$OutOfSchoolChildrenPrimaryAge = \beta_0 + PctPopUnder14 \times \beta_1 + PrePrimaryEnrollment \times \beta_2$$

Variable	Estimate	Std.Error	t.value	pr.t
PctPopUnder14	82430.7	33950.8	2.43	0.01670
PrePrimaryEnrollment	-20136.8	6045.2	-3.33	0.00118

This model was the best estimator of *OutOfSchoolChildrenPrimaryAge* among our models. We were able to explain a small, but considerable portion of the total variance of the response variable while also controlling for difference between countries and over time. Hopefully, as consistency of data collection for WDI indicators continues to improve, these methods will be able to provide more powerful tools for making predictions and inferences in the realm of education and development.

Example: Internet Usage

We were interested in whether our observations regarding health and education indictors are also reflected in technology. We began by plotting the relationship between Internet Users per 100 people and GDP per capita. We chose this predictor because of its highly available data points and intuitive association with technical development. We focused our attention on countries observed during 2005 - 2014. The straight line was fit by OLS. The curvy fit was produced by the lowess method. Evidently, the relationship is not linear.



Now we inspected the variance in Internet Usage by country and year. We immediately observe that overall internet usage grows overtime. There is significant variance between countries.

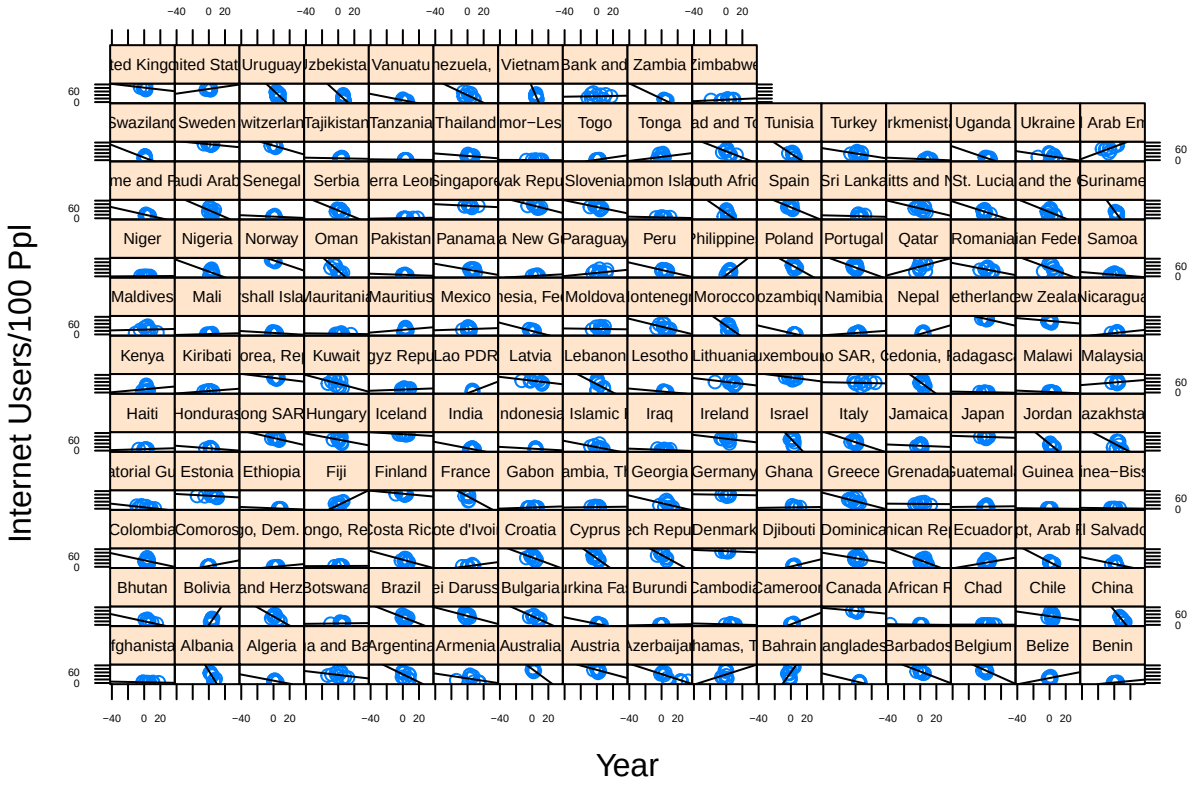
Then we estimate the "within" variance by fitting a simple regression:

$$InternetUsers = \beta_0 + Country \times \beta_1$$

The r-squared for this model is 0.87, and the F-statistic is extremely significant, indicating the grouping by country is an important component of the data.

Next, we used the *psychometric* package to compute the ICC, which indicated that 86% of the total variance in *InternetUsers* was due to the fixed effects. These results confirmed the deviation in Internet Usage by country.

We then plotted the relationship between *InternetUsers* and *GDPperCapita* by country, revealing considerable heterogeneity by country, and within-country variation in *GDPperCapita*:



The ICC of country as a predictor for *GDPperCapita* tells us that that 20% of the variation in GDP is between country, so GDP is a time-variant attribute of countries. Evidently, we are working with multiple subjects (countries) observed over a time series. The cross-sectional information reflected the differences between subjects. The time-series (or within-subject information) reflected in the changes within subjects over time.

OLS Regression

We began with a simple OLS regression fit. *GDPperCapita* was significant in this model, but the result was oddly counterintuitive: Increase in *GDPperCapita* decreases *InternetUsage*:

Variable	Estimate	Std.Error	t.value	pr.t
Intercept	35.37	0.76	46.65	<2e-16
GDPperCapita	-1.26	0.16	-8.17	6.08e-16

This model is also a poor fit with with an r-squared value of 0.04 relative to the 0.87 obtained when we use just country as predictors. The estimates of OLS here are subject to omitted variable bias.

LSDV Regression

Considering the heterogeneity across countries, we next fit a Least Squares dummy variable model. Each component of the country factory is absorbing the effects particular to each country. The adjusted r-squared vastly improved, to 0.94 and *GDPperCapita* became more significant in this model. In exchange, the dummy variables consumed additional 169 degrees of freedom.

Panel Regression

Fixed Effects

With panel regression, we hope to control for omitted variables that differ between cases, but are constant over time, or omitted variables that vary over time, but are constant between subjects. We first run the panel estimators with country fixed effects. This model had an adjusted r-squared value of 0.040 and an F-statistic of 70.38.

Variable	Estimate	Std.Error	t.value	pr.t
GDPperCapita	-0.58	0.07	-8.39	<2e-16

Random Effects

We can compare how the random effects model performs with the data. It appears that the coefficient for *GDPperCapita* is virtually the same as in the fixed effects model. This model had an adjusted r-squared value of 0.042 and an F-statistic of 74.48.

Variable	Estimate	Std.Error	t.value	pr.t
Intercept	33.79	1.95	17.34	<2e-16
GDPperCapita	-0.59	0.07	-8.63	<2e-16

Hausman Test

The Hausman test, however, lead us to reject that the fixed and random effects estimators are giving similar answers. The p-value is so small, 2.253e-06, that we would prefer the fixed effects over the random effects estimate. We also added time-fixed effect to the panel data. The significant increase in the r-squared Lagrange multiplier test to 0.59 indicates the time-fixed effects estimators are preferred.

```
##
##  Hausman Test
##
## data:  IT.NET.USER.P2 ~ NY.GDP.PCAP.KD.ZG
## chisq = 22.3663, df = 1, p-value = 2.253e-06
## alternative hypothesis: one model is inconsistent
```

Cross-sectional Dependence

To test for cross-sectional dependence, we examined if our time series had residuals across countries that are correlated and thus lead to biased results. We used the Breusch-Pagan test against heteroskedasticity and found that in this case, there is cross-sectional dependence.

```
## Warning in pf(stat, df1, df2, lower.tail = FALSE): NaNs produced
```

```
##
```

```
##  F test for individual effects
##
## data:  IT.NET.USER.P2 ~ NY.GDP.PCAP.KD.ZG
## F = 110.1244, df1 = -9, df2 = 1529, p-value = NA
## alternative hypothesis: significant effects
```

Serial Correlation

Last but not least, we tested for serial correlation. The p-value is so small that we rejected the null hypothesis and concluded there is serial correlation.

Heterogeneity in intercept / slopes

Our experimentation earlier suggests that there is considerable heterogeneity in the country-level effects. We used the Breusch-Pagaon test for heteroscedasticity. The test had a very small p-value, indicating the presence of heteroscedasticity. We attempted a model in which intercepts and slopes vary across countries, assuming the effect of GDP is consistent across countries. The model fits better, with a separate slope and intercept to the data from each country The adjusted r-squared value was 0.87 with an F-statistic of 35.67. The combination of country and GDP consumed additional 338 degrees of freedom.

```
##
##  Breusch-Pagan test
##
## data:  m2
## BP = 1149.264, df = 170, p-value < 2.2e-16
```

PGLS Model

Our tests for time fixed effects and auto correlation suggested the need for GLS. Our Panel GLS model returned an r-squared of 0.873. Although still significant, the effect for GDP greatly decreased compared to the Least Squares estimates. This is by far our most powerful explanatory model, balancing r-squared value and degrees of freedom. Meanwhile, our experimentation has shown that technology development trends, too, vary considerably across countries and time frames. There may not be "one model that fits all".

Variable	Estimate	Std.Error	t.value	pr.t
GDPperCapita	-0.08	0.02	-4	6.21e-05