What Makes a Country Wealthier?

Exploring the determinants of GDP per capita

Kevin Chen '15 Varun Sharma '16
Jean-Luc Etienne '15

Abstract

The determinants of a nation's gross domestic product (GDP) have been explored to a great extent. Indeed, given that GDP is a function of a nation's consumption, investment, government spending, exports, and imports, the standard Keynesian model "predicts" GDP nearly perfectly, and quite uninterestingly. We present four core models that try to explain GDP per capita without relying on the literal components of Keynesian aggregate demand. Specifically, they proxy for measures of climate, technological advancement, urbanization and urban welfare, and social development. We find that...

1 Introduction

What makes one country wealthier, or more well off, than another? In Keynesian theory, GDP can be reduced to its literal components:

$$\text{GDP} = C + I + G + (X - M)$$

where C is household expenditure, I is investment, G is government spending, X is exports, and M is imports.

However, regressing GDP on these components would be pointless as it would confirm nothing new, and uninteresting as it would not be an appropriate proxy for a country's standard of living. Instead, we use a more telling response and more interesting indicators.

1.1 Why "per-capita" GDP?

GDP is not a suitable proxy for standards of living. For proof of this, look no further than China. There, GDP has skyrocketed while GDP per capita remains one of the world's lowest. Although GDP per capita may not be a perfect metric for living standards – since not all citizens benefit from a country's increased production – it has been argued that GDP per capita tends to move with living standards. In addition, most countries in the world frequently publish GDP per capita data. Wide access to data, as well as the fact that per-capita GDP is measured with a relatively consistent definition,

makes this a suitable proxy for a country's standards of living.

1.2 Research Question

Our models are inspired from four simple buzzwords: *climate change*, *cities*, *computers*, and *"comfort"* (used somewhat euphemistically). These models beg the following question: Are people better off in greener, more urbanized, more technologically advanced, and more socially developed countries?

For our climate change model, we use metrics for energy consumption, pollution, and paved roads – the idea being that firm production rises with energy consumption, that pollution spikes as a result, and that firms are more likely to produce if they face less congestion when transporting their goods.

For our urbanization model, we use proxies for urban welfare and urbanization level. Cities provide benefits of agglomeration for both firms and workers. In a city, a worker is more likely to match with a firm, to find employment, and to spread knowledge between firms. Similarly, when firms cluster together in a city, they experience greater knowledge spillovers, lower proximity from their buyers, and thus lower transportation costs. In addition, we hypothesize that wealthier urban populations will experience greater productivity and thus lead to greater national output.

Next, for our technology model, we use metrics for high-tech exports, and research and development (R&D). We hypothesize that these metrics are directly correlated with innovation and productivity, and that they rise

with GDP as a result.

Finally, for our social development model, we look at the size of a country's refugee population, the preponderance of child labor, literacy rates, and life expectancy. We hypothesize that literacy rates and life expectancy are positively correlated with GDP, as countries that are healthier and more educated tend to have healthier economies. We also hypothesize that refugee population and child labor are negatively correlated with national output, as they are indicative of poverty, poor schooling, and nearby civil strife.

Given these models, we face several challenges. First, while we have provided hypotheses for each indicator's effect on GDP, we have not hypothesized about their effects on per-capita GDP. That is, a country with higher pollution may be more industrial and thus have a higher output, but what does that say about its population size and consequently about its GDP per capita?

Second, we may face multicollinearity in our models. For instance, our social development model includes literacy rates and child labor. The former is a proxy for education, while the latter is often linked directly to education. This is problematic since multicollinearity increases standard error and decreases an indicator's t-statistic, thus making it appear insignificant when in reality it is not.

The structure of this paper is as follows: We proceed through each model, providing exploratory analysis, condition checking, and interpretation of regression output. Exploratory analysis starts with scatterplots, which show

both the amount of data points for an indicator (so we can weed out indicators with bad data coverage) and the linearity between the indicator and the response. If linearity is not present, we try to transform it, usually with logs. We then perform a regression, interpret the coefficients, and check the model conditions to gauge the model's meaningfulness.

While this is a rough algorithm for our thinking process, it is not something we have to rigidly adhere to. Our goal here is not to come up with the "perfect" model. As previously stated, we can use exports and imports to come up with a model with an exceptionally high R^2 value. Our goal is simply to explain GDP per capita in more relatable ways: people are more knowledgeable about their country's climate and general urban, technological and social environment than they are about its level of exports and imports.

2 Data

We collect data from the World Development Indicators, which draws from several internationally recognized sources and boasts 54 years' worth of data for 214 countries across 1334 indicators. We use data for the year 2008.[1]

In compliance with economic convention, and more importantly, because it makes statistical sense, we use log GDP per capita. This choice is shown in Figure 1. Not only is our response normal, but it is missing very few data points, as shown in the heat map in Figure 2.

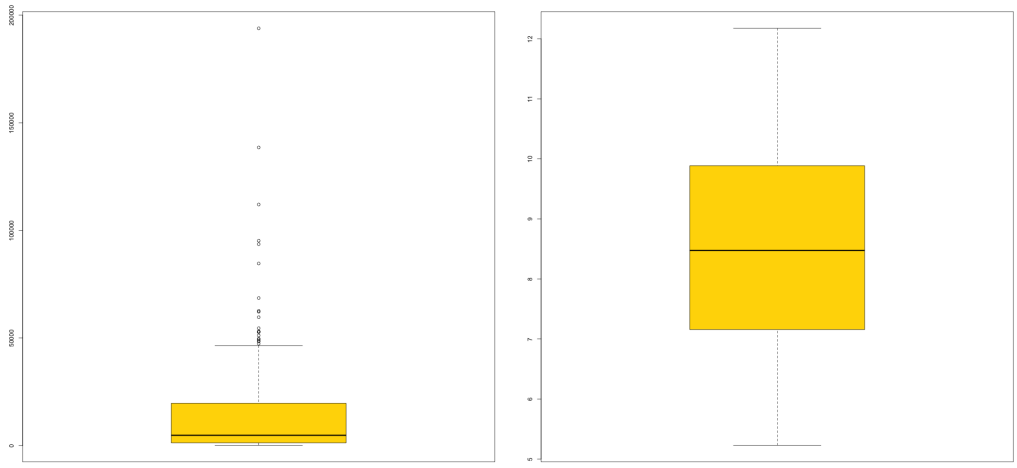

Figure 1: Log GDP per capita is more normal than its unlogged counterpart.

Figure 2: Our response had near-universal coverage.

3 Climate Change Model

3.1 Indicators

How do climate-related indicators affect a country's wealth? To answer this, we look at pollution and energy consumption. Intuitively, there is a correlation between those two: the more energy a country consumes, the more it pollutes. We proceed with caution, wary of multicollinearity.

For our pollution metrics, we choose emissions (measured in kilotons) consisting of three basic chemical compounds: nitrous oxide, carbon dioxide, and methane. Nitrous oxide emissions result from agricultural biomass burning, industrial activity, and livestock management. Carbon dioxide results

from burning fossil fuels. Methane results from agriculture and industry.

Our proxies for energy consumption consist of electric consumption per capita (in kilowatt-hours) and the percentage of total roads that are paved. The former measures the production of power plants (minus transportation costs), while the latter might directly account for those transportation costs: that is, a larger network of paved roads may reduce the time it takes to transport goods and thus energy consumption.

When we run scatterplots of each of the pollution metrics against the response, none resemble anything close to a linear relationship. When we apply a log transformation to each pollution metric, however, all of them exhibit linearity. This correction is illustrated in Figure 3.

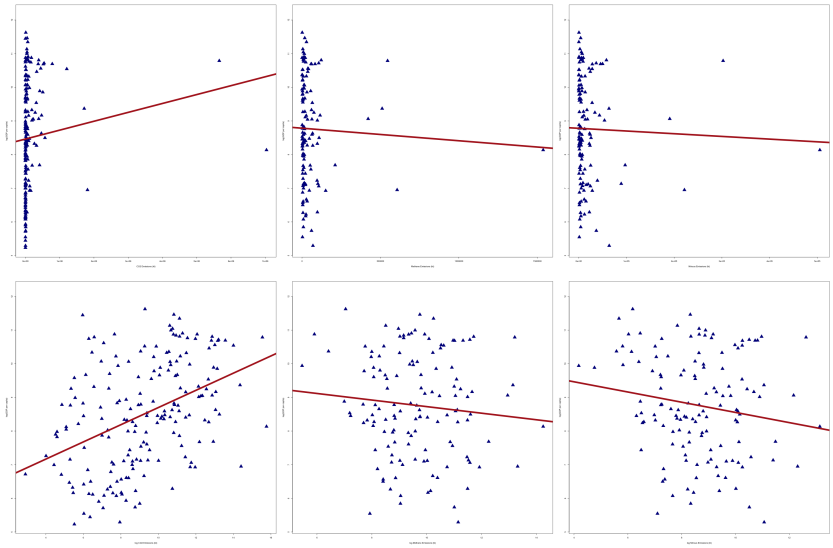

Figure 3: Left to right: CO2, methane, nitrous. Log transformations on bottom row.

Next we run scatterplots of log GDP per capita versus each our proxies

for energy consumption, as shown in the top row of Figure 4. Immediately we see two problems. First, electricity consumption versus log GDP per capita does not exhibit linearity. Second, the number of data points for the paved roads indicator is minimal. A log transformation on each indicator is shown in the bottom row. The log of electricity consumption exhibits a more linear relationship with the response. However, paved roads still has just as few data points.

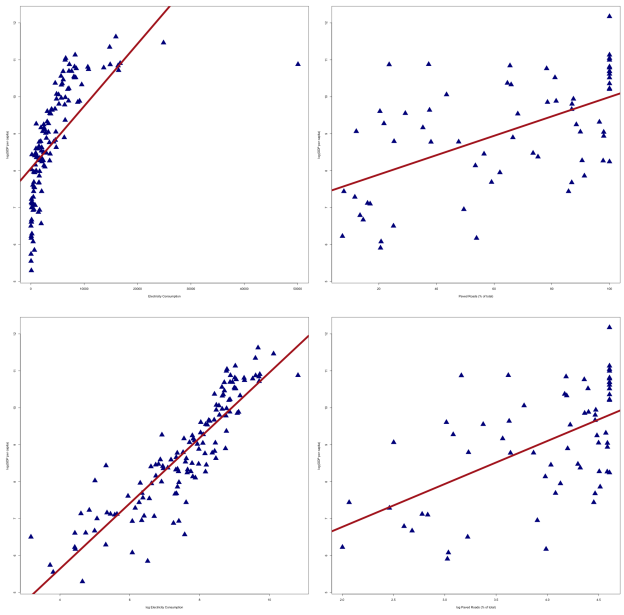

Figure 4: Left to right: electricity consumption, paved roads. Log transformations on bottom row.

3.2 Regression

Figure 5 shows our variations of the climate model. First, we run the full model that includes every indicator. Of these, the log of electricity consumption is the only one to show up as significant. We see a positive correlation: the more energy power plants consume in a country, the higher that country's log per-capita GDP will be. Specifically, for the first three models, a 1% increase in electricity consumption is associated with 0.97% increase in GDP per capita, give or take one basis point.

	(1) Model A	(2) Model B	(3) Model C	(4) Model D
paved_roads	0.00262	0.00166	0.00169	
	(0.79)	(0.56)	(0.59)	
log_electr~s	0.984***	0.961***	0.961***	0.291
	(10.10)	(13.80)	(13.97)	(1.37)
log_co2	-0.0488	0.00220		-0.455**
	(-0.38)	(0.05)		(-2.65)
log_nitrous	0.1000			
	(0.82)			
log_methane	-0.0662			
	(-0.40)			
co2_elec				0.0613**
				(2.82)
_cons	1.465	1.380*	1.399**	6.433***
	(1.69)	(2.23)	(2.89)	(3.96)
N	58	59	59	132

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 5: N increases when we remove paved roads.

However, Model 4 tells a different story. We remove the paved roads

indicator – as it is causing too many values to be deleted – and regress on the interaction between log CO2 and log electricity consumption. We find that the log of electricity consumption is no longer significant. Instead, log CO2 and the interaction term are significant.

3.3 Condition Checking

As with any linear regression model, we provide exploratory analysis of our assumptions: linearity, independence, zero mean, constant variance, and normality.

We confirm (1) linearity above with scatterplots, (2) zero error mean by our use of ordinary least squares linear regression, and (3) independence by the way data points were collected – that is, measuring a nation's statistics does not effect the statistics of it or any other nation. We apply these assumptions to the other three models.

Figure 6 provides a way for us to assess whether our model violates our assumptions of equal variance and normal errors. On the left, we see a scatterplot of residuals versus fitted values, which shows that the model is fairly homoskedastic. On the right, we see the errors are not quite normal, as there is a long left tail. For future work, we might try nonparametric tests, as they do not require a normal error distribution.

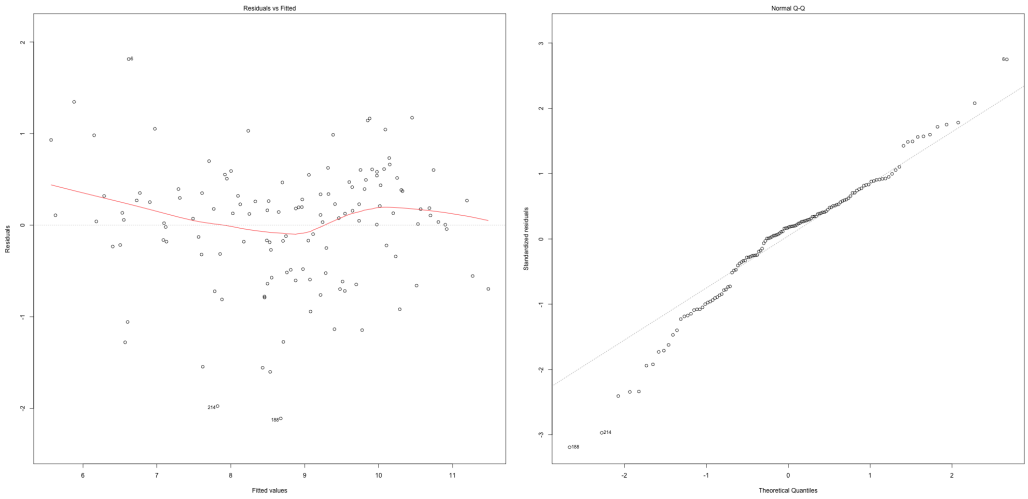

Figure 6: Residuals vs fitted and Q-Q plot

	(1) Model A	(2) Model B	(3) Model C
urb_sanita~n	0.0278***		0.0315***
	(8.52)		(10.95)
urb_popula~n	0.0317***	0.0424***	0.0329***
	(10.18)	(12.65)	(10.75)
urb_water	0.0217*	0.0609***	
	(2.32)	(6.30)	
_cons	2.449**	0.373	4.141***
	(3.26)	(0.44)	(20.85)
N	179	183	181

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 7: We drop urban water in the final model, due to its lack of linearity.

4 Cities Model

4.1 Indicators

4.2 Regression

4.3 Condition Checking

5 Computers Model

5.1 Indicators

5.2 Regression

5.3 Condition Checking

6 Social Development Model

	(1) Model A	(2) Model B	(3) Model C
log_research	0.324		
	(1.83)		
log_high_t~h	0.179	0.0904	-0.102
	(1.38)	(1.23)	(-1.13)
log_journals	0.212**	0.220***	0.108*
	(2.97)	(5.32)	(2.08)
tech_jour			0.0661***
			(3.41)
_cons	7.639***	7.407***	7.620***
	(11.71)	(34.58)	(35.32)
N	79	143	143

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

8 Conclusion

summarize results, describe limitations and how they might be addressed

Notes

[1]For modularity, we built a script that created a data frame for each year of data downloaded, and then recast each frame from wide to long format.

	(1) Model A	(2) Model B	(3) Model C
child_labor	-0.0426		
	(-2.19)		
literacy		0.0462*	
		(2.48)	
life_expect			0.133***
			(16.19)
refugee_pop			-0.000000429
			(-1.48)
_cons	8.381***	4.421*	-0.441
	(15.09)	(2.74)	(-0.79)
N	12	25	180

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 8: Sufficently large n required removing child labor and literacy.