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 change, 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.

For each model, we provide 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. If we cannot achieve linearity, we remove the indicator from our model.[1] 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.

The structure of this paper is as follows. Section 2 describes the data source. The following sections describe indivual models; that is, they try to see how well a group of indicators explains GDP per capita. Section 3 gives an account of our climate change model; Section 4, our urbanization model; Section 5, our technology model; and Section 6, our social development model. Section 7 concludes.

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.

[1]Due to a model's tendency to change, we often provide a side-by-side comparison of all its forms. The table includes coefficients and t-statistics for each indicator, as well as one, two, or three asterisks for the α levels of 0.05, 0.01, and 0.001, respectively.

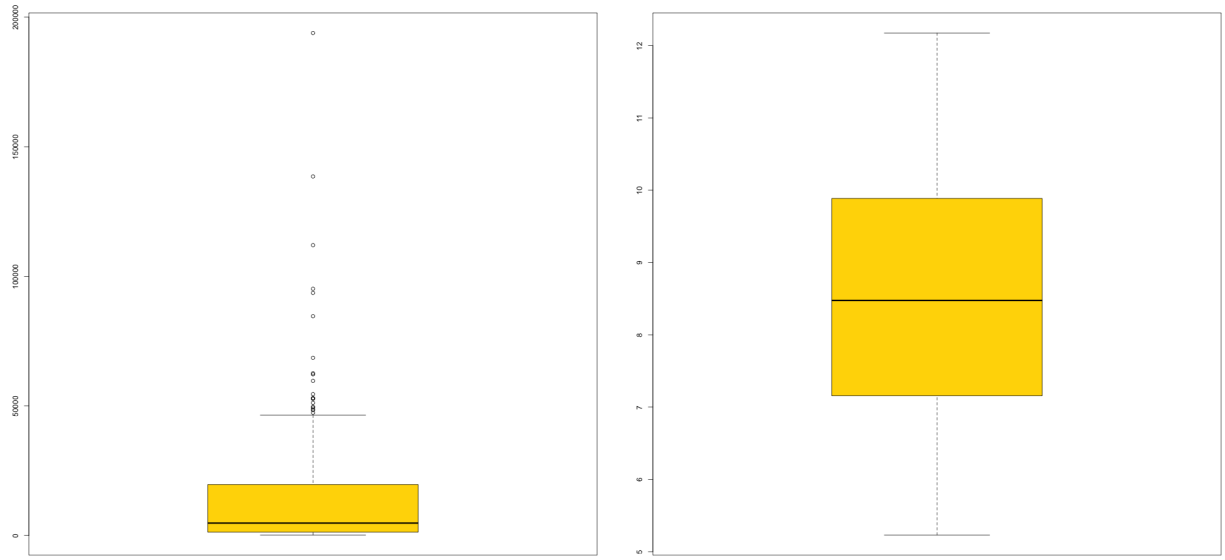

Figure 1: Log GDP per capita is more normal than its unlogged counterpart.

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.

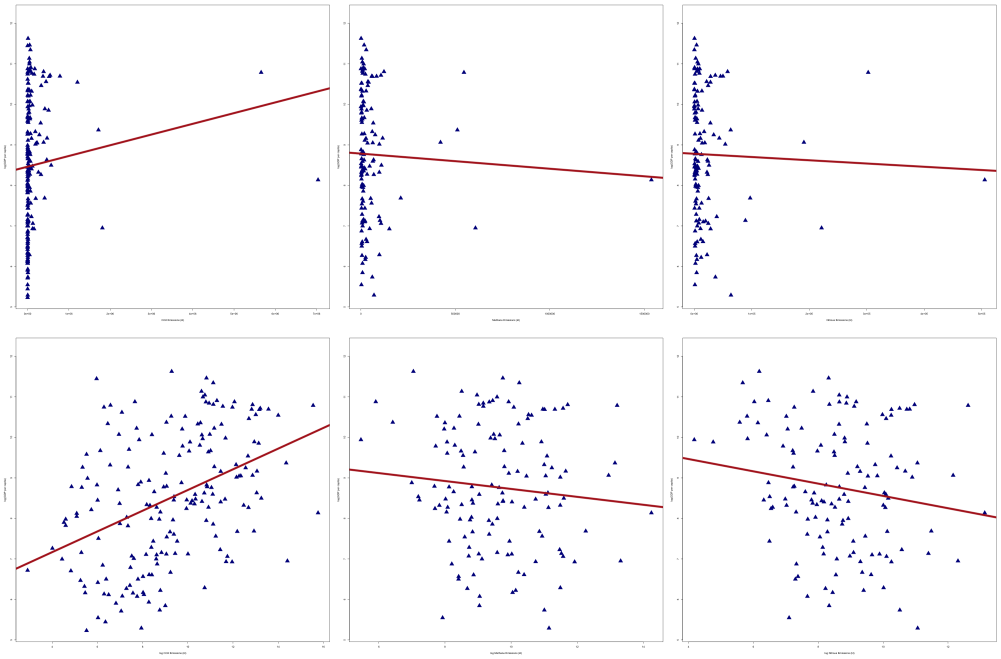

Figure 3: Left to right: CO2, methane, nitrous. Log transformations on bottom row.

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.

3.2 Regression

Table 1 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 models 1 and 3, a 1% increase in electricity consumption is associated with 0.97% increase in GDP per capita, give or take one basis point.

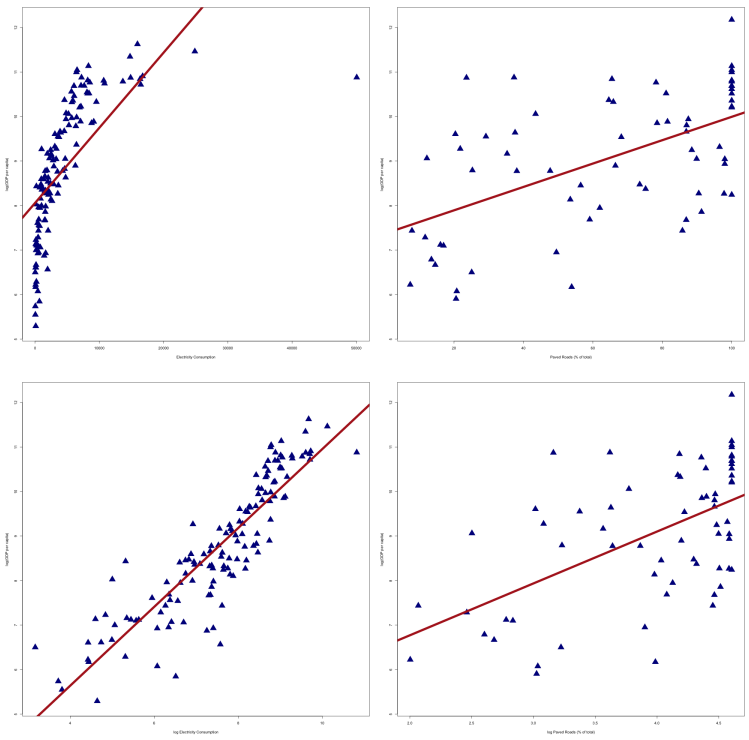

Figure 4: Left to right: electricity consumption, paved roads. Log transformations on bottom row.

Table 1: N increases when we remove paved roads.

	Dependent variable:			
	log.GDP.per.capita			
	(1)	(2)	(3)	(4)
paved.roads	0.003	0.001	−0.001	
	(0.003)	(0.003)	(0.016)	
log.electricity.consumption	0.984***	0.411	0.962***	0.291
	(0.098)	(0.313)	(0.071)	(0.212)
paved.roads:log.co2			0.0003	
			(0.002)	
log.co2	−0.049	−0.454*	−0.017	−0.455***
	(0.127)	(0.257)	(0.106)	(0.172)
log.nitrous	0.100			
	(0.122)			
log.methane	−0.066			
	(0.165)			
log.electricity.consumption:log.co2		0.058*		0.061***
		(0.032)		(0.022)
Constant	1.465*	5.736**	1.557	6.433***
	(0.865)	(2.491)	(1.105)	(1.626)
Observations	58	59	59	132
R^2	0.842	0.840	0.831	0.806
Adjusted R^2	0.827	0.829	0.818	0.802
Residual Std. Error	0.591 (df = 52)	0.584 (df = 54)	0.601 (df = 54)	0.668 (df = 128)
F Statistic	55.543*** (df = 5; 52)	71.067*** (df = 4; 54)	66.321*** (df = 4; 54)	177.733*** (df = 3; 128)
Note:				*p<0.1; **p<0.05; ***p<0.01

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 5 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.

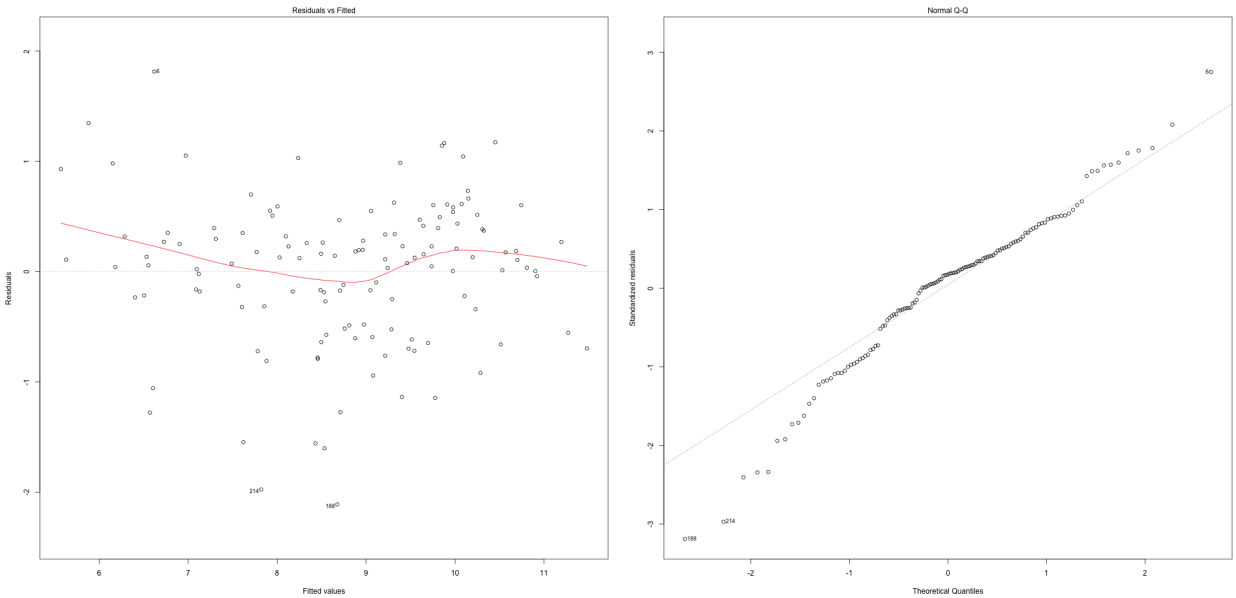

Figure 5: Residuals vs fitted and Q-Q plot

4 Cities Model

4.1 Indicators

How well does a country's urban environment explain its wealth? To answer this, we look at the size of a country's urban environment, as well as the welfare of that environment. For size, we use the percentage of population living in urban environments. For welfare, we use the percentage of the urban population with access to clean water. Figure 6 shows scatterplots of log GDP per capita against our *population* and *water* indicators.

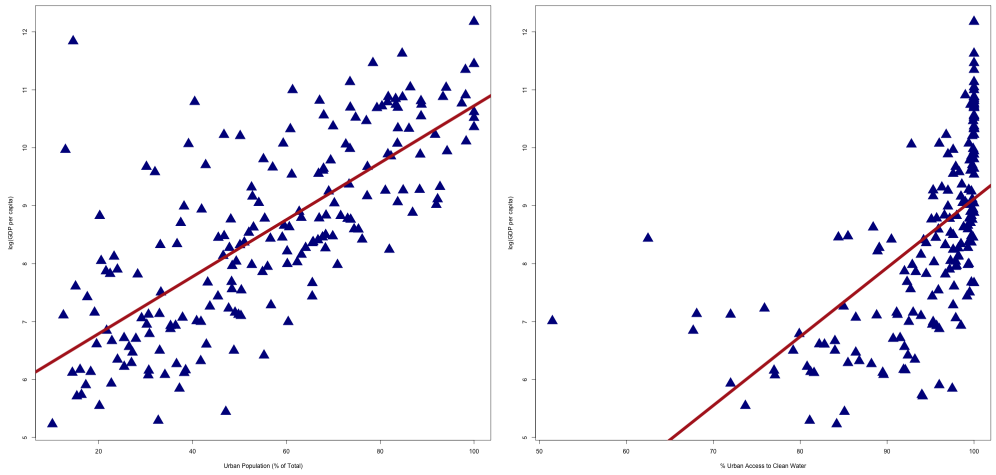

Figure 6: Left to right: Population, water.

We lack a linear relationship between the response and the percentage of urban population with access to water. We tried to remedy this defect with several transformations, but to no avail. Consequently, we replace our urban welfare predictor with the percentage of the urban population with access to sanitation, which primarily entails access to latrines. This new indicator exhibited a much more linear relationship with log GDP per capita, as seen in Figure 7.

4.2 Regression

Table 2 shows the regressions of various phases of our urban model. The first model includes our size metric (urban population) as well as both of our proxies for urban welfare (urban sanitation and urban water). Model 2 regresses log GDP per capita on urban population and urban water, whereas Model 3 regresses on urban population and urban sanitation.

Table 2: We drop urban water in the final model, due to its lack of linearity.

	Dependent variable:		
	log.GDP.per.capita		
	(1)	(2)	(3)
urban.population	0.032***	0.042***	0.033***
	(0.003)	(0.003)	(0.003)
urban.sanitation	0.028***		0.032***
	(0.003)		(0.003)
urban.water	0.022**	0.061***	
	(0.009)	(0.010)	
Constant	2.449***	0.373	4.141***
	(0.752)	(0.841)	(0.199)
Observations	179	183	181
R^2	0.761	0.664	0.756
Adjusted R^2	0.757	0.661	0.753
Residual Std. Error	0.781 (df = 175)	0.922 (df = 180)	0.787 (df = 178)
F Statistic	186.043*** (df = 3; 175)	178.061*** (df = 2; 180)	275.727*** (df = 2; 178)
Note:			*p<0.1; **p<0.05; ***p<0.01

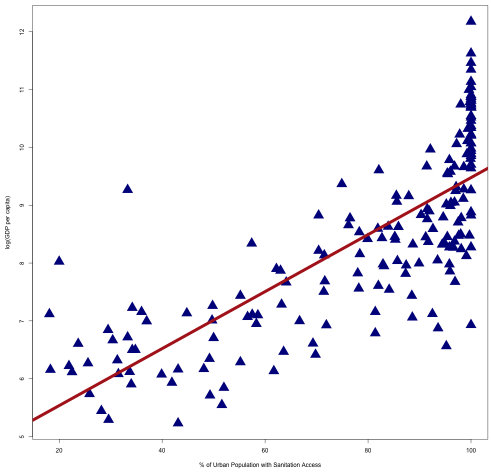

Figure 7: Access to sanitation is a better proxy for urban welfare than access to water.

Every indicators in each of the models appears significant, but we should expect the highest R^2 from the last model, since its welfare proxy (*urban sanitation*) bears a much more linear relationship with the response than the other welfare proxy (*urban water*). Indeed, the R^2 of Model 2 is 0.66, whereas the R^2 of Model 3 is 0.75.

4.3 Condition Checking

Figure 8 allows us to assess our model condition's of equal variance and normal errors. The scatterplot of residuals versus fitted values shows roughly equal variance for all fitted values. The Q-Q plot, on the other hand, is not as promising, as we see a slightly long left tail. If we believe the lack of normality in the errors is a concern, we could perform nonparametric tests, which punt on assumptions about the error distribution.

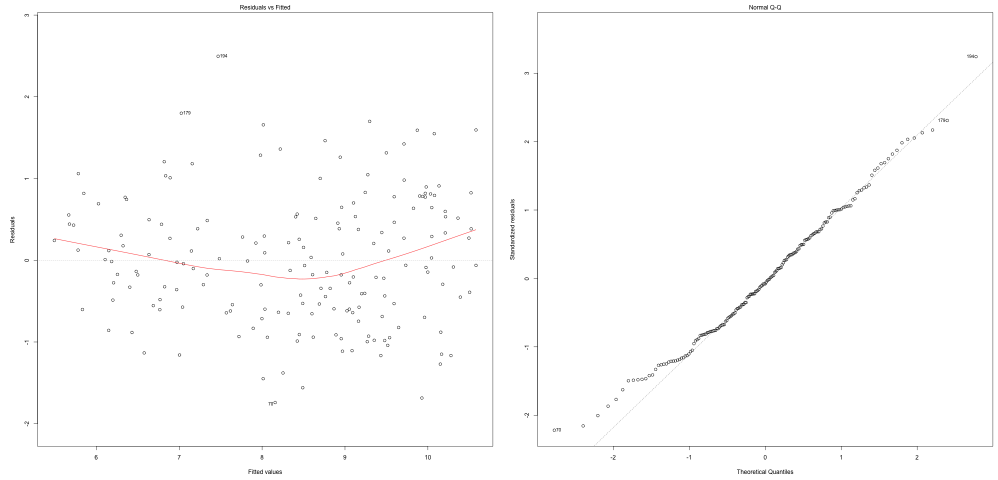

Figure 8: The urban model shows equal variance and (almost) normal errors.

5 Computers Model

5.1 Indicators

How well does a country's level of technology explain its wealth? We examine three indicators: the number of scientific journals and articles produced by a nation, the percent of GDP that goes into research and development (R&D), and the percentage of manufactured exports that are deemed "high-tech." It is worth noting that each of these explanatory variables may be highly correlated

with each other. In other words, the more a country invests in R&D, the more scientific journals it will produce, and vice versa.

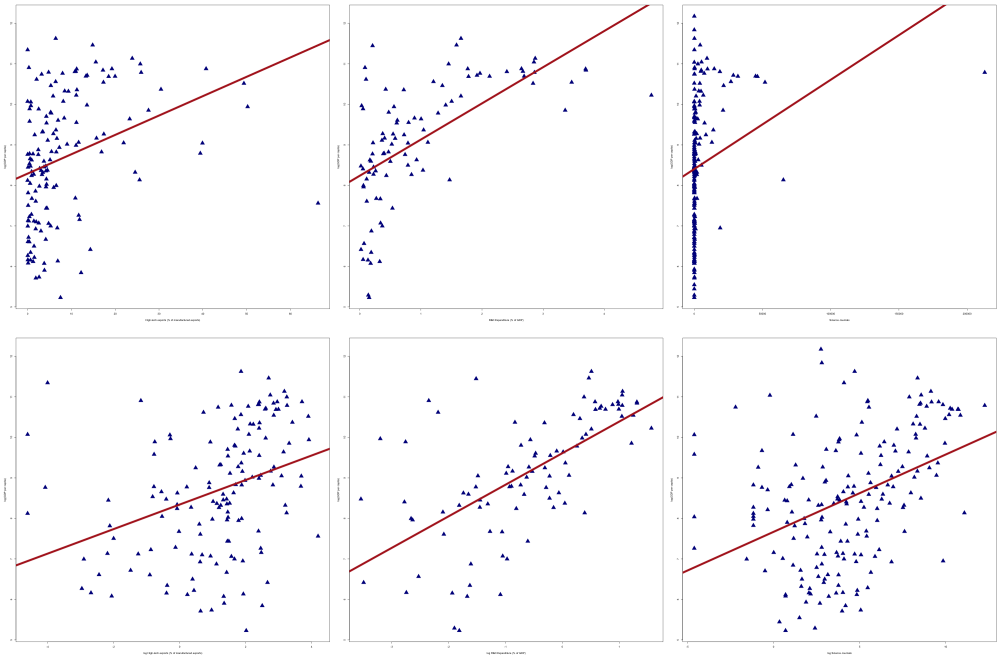

Figure 9: Left to right: exports, R&D, journals. Logs on bottom.

5.2 Regression

Once we apply log transformations to each of our three indicators, we see strong linear relationships, shown in Figure 9.

Table 3 shows various forms of our science and technology model. We use the logarithms of our World Bank indicators, as they are stronger predictors of log GDP per capita.

In addition, we include every possible interaction term for two reasons. First, we do so because we can afford to. With only $K = 3$ explanatory variables, we can show all $6 = K(K-1)/2$ pairwise combinations. Second, and more importantly, we suspect that the effect of one explanatory variable on log GDP per capita depends on the level of another covariate. For instance, research may have less impact on GDP for lower levels of high-tech exports. This is intuitive since, in our model, research is only as good as the fruits borne by it. In short, we justify our use of interaction terms with the following question: What's the point of science journals and R&D if you can't make things?

Table 3 answers precisely that question. Indeed, there is no "point" – in the sense of output and GDP – to the science community if it doesn't produce anything. We know this because $\ln(\text{research}) \times \ln(\text{high.tech})$ and $\ln(\text{high.tech}) \times \ln(\text{journals})$ are significant, but $\ln(\text{research}) \times \ln(\text{journals})$ is not.

Table 3: Taking out *research* drastically increases the number of observations.

	Dependent variable:				
	log.GDP.per.capita				
	(1)	(2)	(3)	(4)	(5)
log.research	0.324*	0.576**	−0.076		−0.758**
	(0.178)	(0.277)	(0.225)		(0.339)
log.journals	0.212***	0.206***		0.108**	0.661***
	(0.071)	(0.070)		(0.052)	(0.151)
log.high.tech:log.journals				0.066***	−0.237***
				(0.019)	(0.063)
log.high.tech	0.179		0.512***	−0.102	2.115***
	(0.129)		(0.161)	(0.091)	(0.429)
log.research:log.journals		−0.018			0.021
		(0.041)			(0.045)
log.research:log.high.tech			0.367***		0.544***
			(0.093)		(0.098)
Constant	7.638***	8.054***	8.301***	7.620***	3.785***
	(0.652)	(0.559)	(0.408)	(0.216)	(1.057)
Observations	79	88	81	143	79
R^2	0.499	0.521	0.504	0.326	0.655
Adjusted R^2	0.479	0.504	0.484	0.311	0.626
Residual Std. Error	1.050 (df = 75)	1.067 (df = 84)	1.042 (df = 77)	1.282 (df = 139)	0.890 (df = 72)
F Statistic	24.942*** (df = 3; 75)	30.418*** (df = 3; 84)	26.041*** (df = 3; 77)	22.394*** (df = 3; 139)	22.737*** (df = 6; 72)

Note: *p<0.1; **p<0.05; ***p<0.01

5.3 Condition Checking

Figure 10 shows the regression conditions for the following model:

$$\ln(\text{GDP per capita}) = \beta_0 + \beta_1 \cdot \ln(\text{research}) + \beta_2 \cdot \ln(\text{exports}) + \beta_3 \cdot \ln(\text{journals}) +$$
$$\beta_4 \cdot \ln(\text{research}) \times \ln(\text{journals}) +$$
$$\beta_5 \cdot \ln(\text{exports}) \times \ln(\text{research}) +$$
$$\beta_6 \cdot \ln(\text{exports}) \times \ln(\text{journals})$$

From the plot of residuals versus fitted values, we can glean that there is fairly equal variance. From the Q-Q plot, we see that the errors are roughly normal, with some deviation on the left tail.

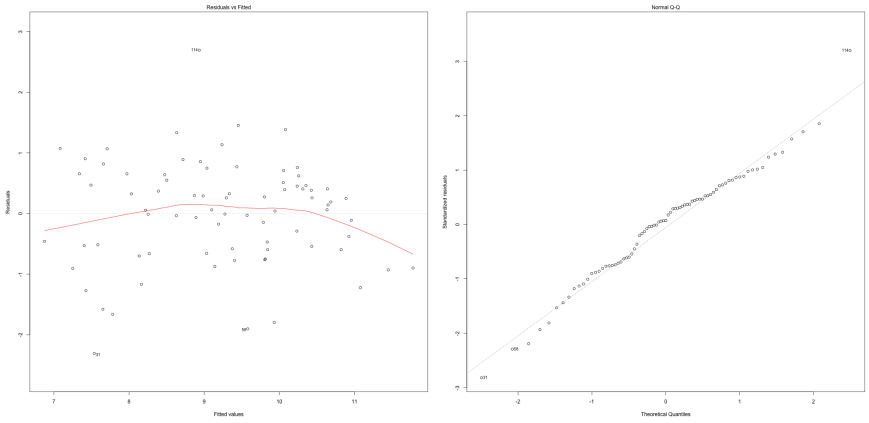

Figure 10: Left to right: Residuals vs fitted values, Q-Q plot.

6 Social Development Model

6.1 Indicators

How does a country's level of social development affect its output, and more importantly, its GDP per capita? We look at four primary indicators: the unemployment rate for children ages 7-14, the life expectancy at birth (in years) for males, the refugee population, and the literacy rate for adults aged 15 and above. These indicators measure a country's economic, health, political, and educational institutions, respectively. Intuition tells us they are positively correlated with national output. That is, we expect GDP per capita to rise with lower unemployment rates, higher life

expectancy, a lower refugee population, and higher levels of education.

6.2 Regression

Table 4 shows our social development model. Model 1 and model 2 simply show that child labor and literacy, though significant, do not have enough observations. For this reason, we focus on model 3, which shows that ln(life.expect) and ln(refugees) have an extremely statistically significant effect on ln(GDP.per.capita).

Table 4: Sufficently large n acquired from removing child labor and literacy.

	Dependent variable:		
	log.GDP.per.capita		
	(1)	(2)	(3)
log.child.labor	-0.968^{**}		
	(0.314)		
log.literacy		3.234^{**}	
		(1.387)	
log.life.expect			6.981^{***}
			(0.534)
log.refugees			-0.136^{***}
			(0.023)
Constant	10.128^{***}	-5.968	-19.888^{***}
	(0.934)	(6.156)	(2.303)
Observations	12	25	180
R^2	0.487	0.191	0.633
Adjusted R^2	0.436	0.156	0.628
Residual Std. Error	0.951 (df = 10)	1.278 (df = 23)	0.968 (df = 177)
F Statistic	9.496^{**} (df = 1; 10)	5.438^{**} (df = 1; 23)	152.374^{***} (df = 2; 177)
Note:			*p<0.1; **p<0.05; ***p<0.01

6.3 Condition Checking

The conditions of model 3 are depicted in Figure 11. From the plot of residuals versus fitted values, we see that the variance is not fan-shaped, but it is also far from ideal. From the model's Q-Q plot,

we see that the error distribution is far from normal, since both tails are long. Thus, our social development model may require a nonparametric test.

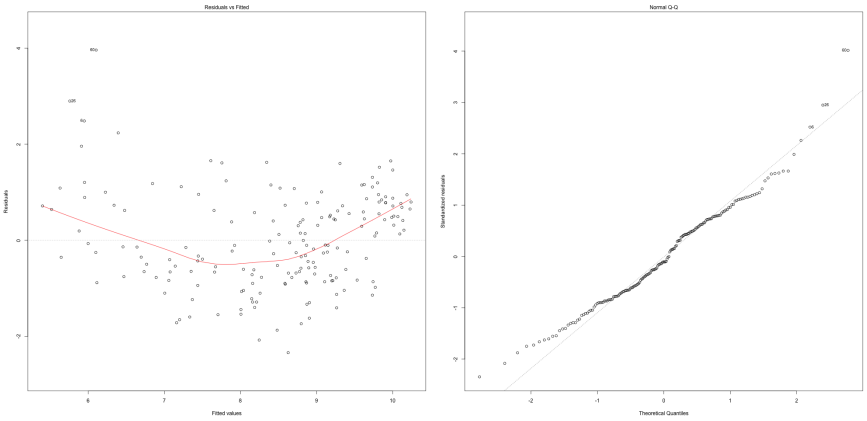

Figure 11: Left to right: Residuals vs fitted values, Q-Q plot.

7 Conclusion

We provide four core models that explain the log of GDP per capita. Each model has very significant indicators and high R^2.

However, there may an inherent limitation to our regressions: reverse causality. In each model we try to explain the effect of x on y, when in reality it could be that y affects x. For instance, a country's GDP per capita may be what allows it to invest more in research and technology.

Even if y does not directly cause x, it may affect x through another lurking variable. We must think about what affects the indicators we use. First, we do not account for a nation's institutions, even though they have a huge impact on just about everything. Wealthier nations can afford to have strengthen their legal and political institutions. As a result, regulations on pollution or on child labor may be more stringent.

Another limitation might be multicollinearity between two or more indicators in a regression. Often, two explanatory variables are literally proxies for the same thing, as was the case with methane emissions and nitrous emissions; they both proxy for pollution. Sometimes one covariate affects another. This occurred in the climate change model, since higher energy consumption almost certainly increases pollution. Additionally, we speculate that it may have occurred in the social

development model: higher literacy rates may rise with productivity, which may eliminate the need for children workers.

In the case of multicollinearity, regressions can fail. In the case of reverse causality, covariates can be correlated with the error terms. There are several methods to fix the problem of biased and inconsistent estimates – such as instrumental variables, decision tree predictor selection, stepwise regressions – but they are all beyond the scope of introductory statistics.

By far, the easiest way to deal with bias is to manually remove one or more of the explanatory variables. Collinear covariates can be weeded on mere suspicion, or based on the variance inflation factor (VIF). For instance, if the VIF for *nitrous emissions* is 9, this means $SE(\beta_{\text{nitrous emissions}})$ is $\sqrt{9} = 3$ times as large as it would be if *nitrous emissions* were not correlated with any other predictors.

Our models are not without their flaws, but we have provided a preliminary investigation into what is required to explain a country's per capita wealth in a robust and unbiased way.

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.