Empirical Economics (2018) Assignment 3

Suggested Solutions – Andrew Proctor

Note: For this assignment, you are asked to perform an analysis of the impact of "quality of government" (QoG) on economic development. Compared to previous assignments, the way you conduct the analysis is much more open-ended. In the solutions that follow, I provide one example of how to answer this question, while making sure to meet the various requirements outlined in the assignment instructions. Rather than compose these answers as a series of standalone responses, I have chosen to structure the assignment solutions as a simplified report much like you would see in an actual research analysis. Both the format and content of your solutions may differ quite a bit from what you see here – that's okay. By design, there is no single correct answer to this assignment. The important thing is to provide careful and thorough answers for the tasks outlined in the assignment.

Introduction

One of the most commonly theorized determinants of economic development of nations is the quality of their government and institutions. This analysis will test this hypothesis using data from the Quality of Government Institute (at the University of Gothenburg) and replication data from "The Colonial Origins of Comparative Development" by Acemoglu et al. Using this data, I first select a suitable proxy for quality of government, as well as suitable control variables to reduce potential omitted variable bias in my analysis. I then test the hypothesis that quality of government affects economic development using the following regression methods: pooled OLS, random-effects and fixed-effects regression. Using these methods, I estimate that better quality of government increases economics development. However, for reasons I will discuss later in the analysis, I do not believe the regressions used in this analysis satisfy the necessary assumptions for valid causal inference.

Data

For this analysis, I use panel data on 150 countries over the period 1991 to 2016. This dataset allows me to explore the effect of quality of government on economic development for the large majority of countries over approximately the last twenty-five years for which there is data. The sample of countries corresponds to those included in Quality of Government dataset with at least ten years of complete observations for the explanatory variables used in my analysis. This ensures that there is adequate variation to conduct fixed-effects regression, which uses within-country variation.

Variables of Analysis

To begin, I choose suitable variables for the main explanatory variable (proxying "quality of government"), as well as control variables and the dependent variable indicating economic development.

Quality of Government

There are a number of possible variables that could be used to proxy quality of government. Some notable choices include the "Control of Corruption estimate," the "Political corruption index," the "ICRG Indicator," and the "Bayesian Corruption Indicator."

All of these are fairly broad measures of quality of government. This broadness is largely positive – as it encompasses several dimensions of quality of government. At the same time, broad measures like these can be somewhat more difficult to interpret, because of their construction as aggregate indices of several underlying variables, many of which are subjective.

For this analysis, I use the "Political corruption index," originally from the Varieties of Democracy Dataset. The "Political corruption index" is measured on a 0-1 scale; with higher corruption implying a value closer to one.

This variable attempts to measure corruption in three branches of government: the executive, legislature, and judiciary. This indicator is relatively appealing for three reasons. First, corruption is a commonly used indicator for quality of government. Second, the dimensions of the variable (while focusing on broad governance), appear somewhat less subjective than alternative broad measures of quality of government. Finally, the variable is available for most countries over a rather long period of time, with data on corruption and GDP reported for the vast majority of countries after 1991.

Dependent Variable

The measure of economic development I choose to adopt is gross domestic product (GDP) per capita, reported in terms of constant dollars purchasing power parity (PPP). I moreover log-transform GDP per capita, so that marginal effects can be interpreted as percent changes in GDP.

Control Variables

In selecting control variables, I choose variables which are intended to satisfy two conditions. First, the variables should be correlated with both GDP and corruption, so that their omission could bias the results. Second, the variables should not be intermediate outcomes of corruption, such that by including them in the analysis, I would be "over-controlling" for part of the effects of corruption on GDP.

The first set of variables that I adopt encompass hypothesized determinants of the institutional structure of the nation (thereby potentially influencing the quality of government). If the determinants of institutions affect GDP in some other way than through their effect on quality of government, then omitting these variables would bias the estimated results. Consequently, I control for the following two types of determinants of institutions: the legal/governmental origins of the country and the climate/disease environment of the country (which is Acemoglu et al hypothesize affected who chose to settle in a given area). All of the institutional determinants controls are from the Acemoglu et al. replication dataset.

For government and legal origins, I include dummy variables for whether the country was a former colony of the British or French and whether the legal system has French legal origin (all as indicator variables). For the climate and disease environment of a country, I include the absolute value of the latitude of the country's capital (thereby measuring distance to equator), mean temperature of the country, the contemporary presence of yellow fever (as an indicator), and the contemporary severity of malaria (reported as an index from 0-1). Note that in the Acemoglu et al data, the yellow fever variable is listed as "yellow fever present today" (the labelling of which I have preserved here). It is unclear when exactly "today" corresponds to – but I interpret this as proxying contemporary presence of yellow fever. Finally, I note that all of the institutional determinant controls are reported as time-invariant measures (meaning that they have just one value that does not change over time).

As a second set of controls, I also include time-varying variables that are intended to capture plausibly exogenous factor endowments. The variables I include are the country's population and natural resources (oil and natural gas). These variables are from the Quality of Government dataset. Oil and natural gas production are each measured in tens of billions of 2014 US dollars, while population is

measured in 100 millions of persons. The population measure is moreover lagged by five years, so that population size is less likely to be affected by the current GDP or the quality of government environment.

Descriptive Statistics and Exploratory Analysis

Summary statistics for all countries over the sample period (1991-2016) are reported in Table 1 below. There are several notable observations from the summary statistics. First, I note that there are 3900 observations of "year" within the dataset – although there are only 26 years of data. This is because the number of observations corresponds to the number of country x year pairs (that is 150 x 26 years). Most other variables also have a fairly small fraction of non-missing observations, except for mean temperature (only 1378 / 3900 non-missing observations).

I also note that oil and natural gas production have very large variances relative to their means, suggestive of many countries having no or minimal energy production, while some have very large production levels. Finally, I note that the pronounced influence of French and British legacies, with 30% of observation from former British colonies and 48% of observations from countries whose legal system originated from the French legal system.

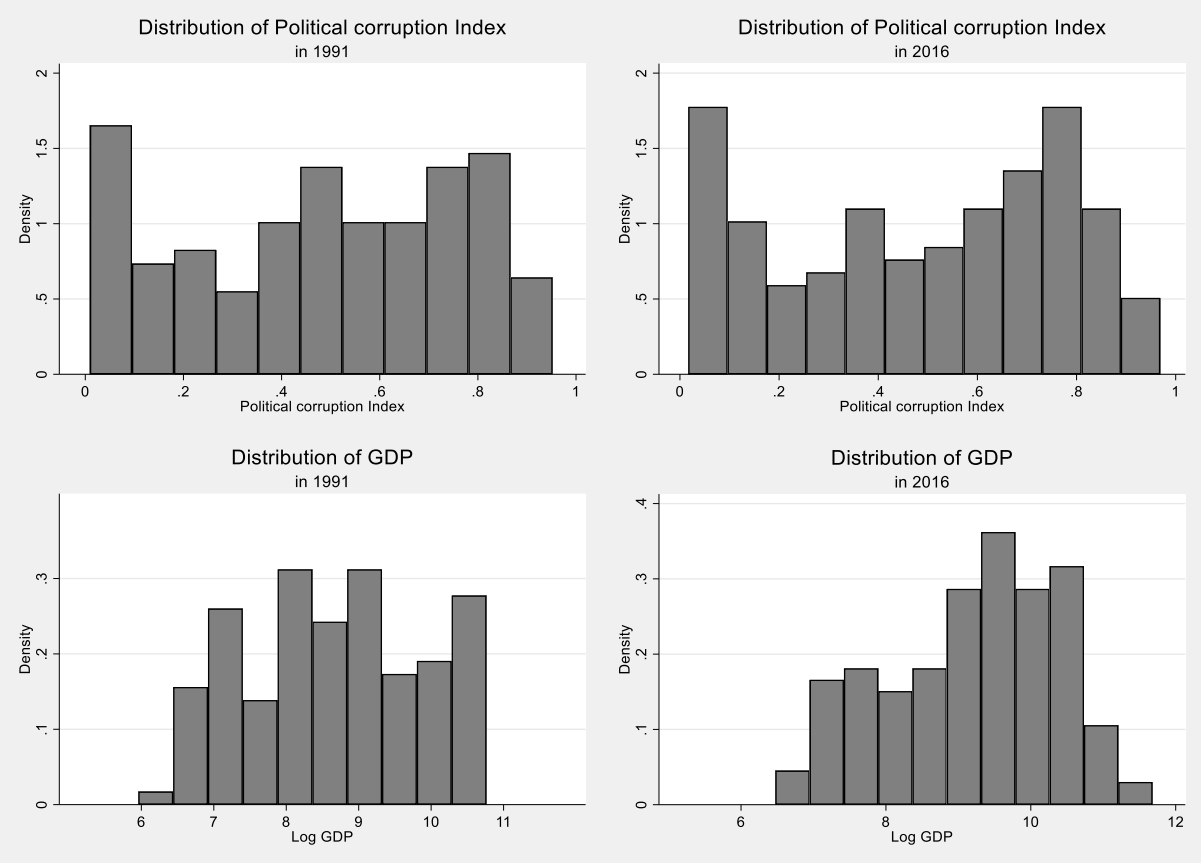
Table 1: Summary Statistics of Key Variables

	Observations	Mean	SD	Min	Max
Year	3900	2003.50	7.50	1991	2016
Log of GDP (2011 Constant Dollars PPP)	3772	8.91	1.23	5.51	11.77
Political corruption index	3856	0.51	0.29	0.01	0.97
Time-Varying Factor Input Controls					
Population (Lagged, 100 millions)	3269	0.26	0.44	0.00	3.21
National oil production (in 2014 dollars, 10 billions)	3557	0.88	3.26	0.00	41.77
National gas production (in 2014 dollars, 10 billions)	3479	0.39	1.69	0.00	29.64
Time-Invariant Determinant of Institutions Controls					
Latitude of capital (absolute value)	3666	30.20	19.20	1.11	72.22
Mean temperature	1378	22.88	5.16	-0.20	29.30
Yellow fever present today	3666	0.48	0.50	0	1
Malaria index in 1994	3614	0.28	0.39	0.00	0.95
Former British colony	3666	0.29	0.45	0	1
Former French colony	3666	0.14	0.35	0	1

Variation over time and across countries

Since each variable is reported across both many countries and many years, it is hard to interpret the source of variation in the summary statistics. To get a clearer picture of variation for the key variables of the analysis (GDP and the corruption quality-of-government measure), I conduct further exploratory analysis for these variables. From the histograms in Figure 1, it is possible to get a clearer sense of what the variation in GDP and political corruptions looks like both across countries and over time. For political corruption, we notice than in the initial period (1991), there is a relatively large share of countries with very low levels of corruption (approximately 0-0.1 corruption index scores), but otherwise the share of countries within a given corruption range does not vary greatly. By 2016, there has been a relative shift out away from moderate levels of corruption (index score of about 0.4-0.6),

Figure 1

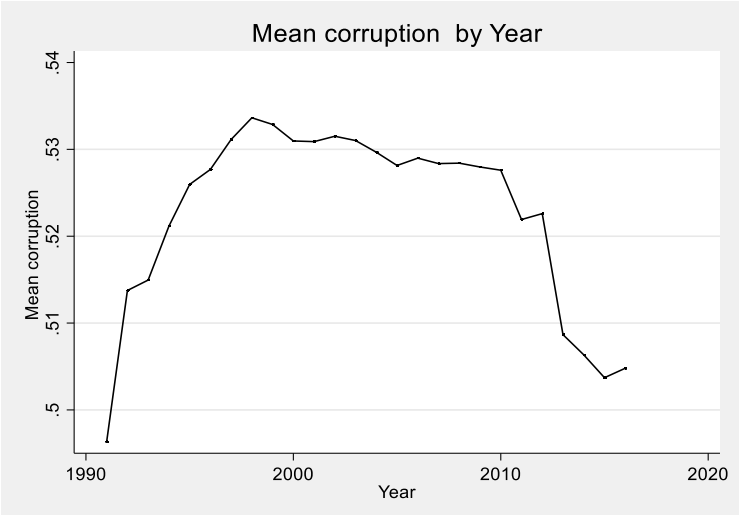


with greater frequency of higher levels of corruption. Data on the corruption index score is not present for every country and every year (as I will discuss further shortly). Importantly, the changes in corruption I note here do not seem to be driven by a different sample of countries from one year to the next, as I obtain similar shifts in corruption when restricting the graphs to only include countries with no missing data.

For GDP, we observe that the distribution has also become more right-skewed in recent years, indicating that fewer countries have very low absolute levels of income and suggesting there has been some convergence in income.

From Figure 2, we see that that the increase in corruption observed from the histograms is largely due to an increase between 1991-2000, with mean corruption rates roughly stabilizing and subsequently declining after the year 2000.

Figure 2



Finally, using the *xtsum* command in Stata, I find that most of the variation in the corruption measure is between the average level of corruption in different countries (standard deviation of .263) rather than within countries over time (standard deviation of .099). This is noteworthy because having too little variation within subjects over time makes inference with

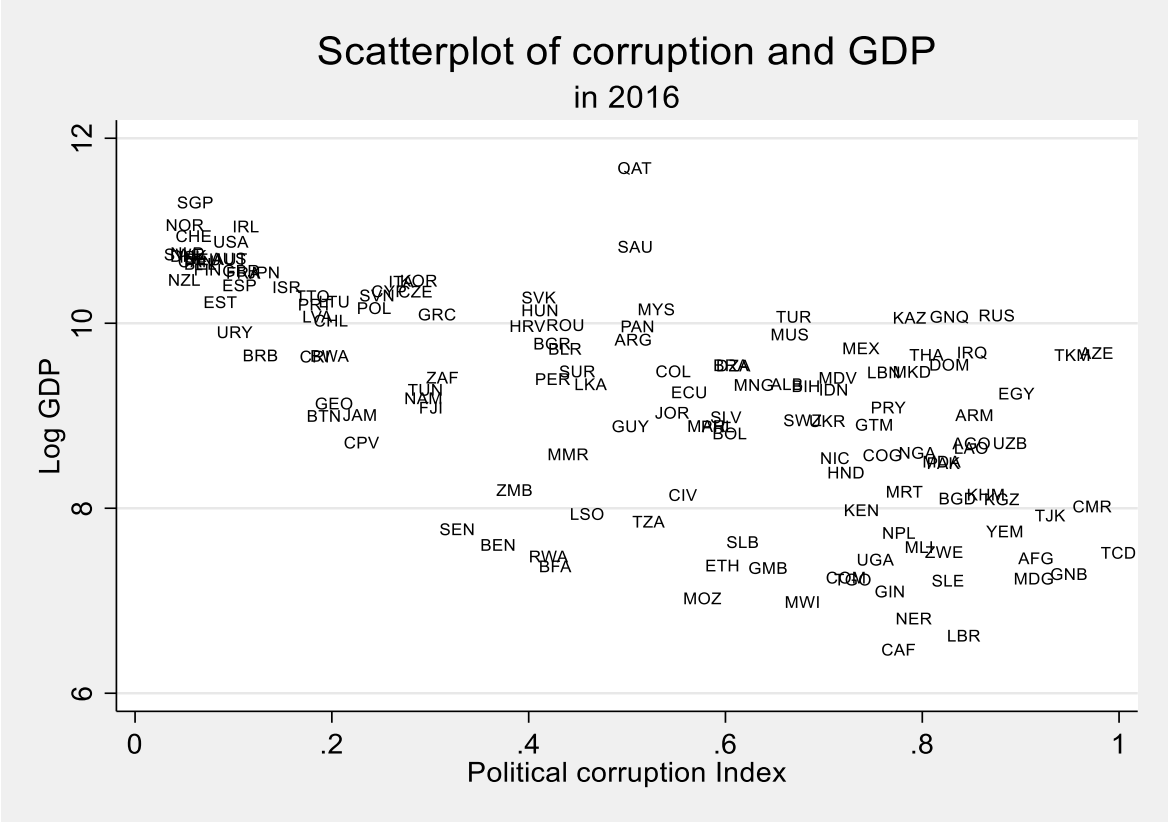
fixed effects regression more difficult by making estimates less precise, since fixed effects regression uses within-subject variation. Because pooled OLS uses all sample variation to produce estimates of explanatory variable parameters, it is considerably more efficient when there is little within-subject variation.[1]

Although the low relative within-country variation in corruption will tend to make fixed effects regression less efficient, we must also consider potential biasedness of a given method. In many cases, fixed effects regression provides more unbiased results since it can control for any time-invariant omitted factors (unliked pooled OLS) and does not impose the same uncorrelatedness assumption between country-specific effects and explanatory variables. However, in this instance, it is unclear which method is superior, since it is unclear what the source of variation in corruption/quality of government is after controlling for long-term determinants of quality of government.

Relationship between GDP and Corruption

By plotting corruption against GDP in Figure 3, we observe that GDP per capita is falling in the corruption index score, as one would expect if corruption is harmful for the economy. Table 3, which lists the countries with the highest or lowest values of corruption, seems to suggest the same correlation. Post-Soviet states, however, tend to have some of the highest levels of corruption while having relatively better GDP ranks.

Figure 3



[1] Additionally, if we believe that conditionally exogenous sources of variation in quality are stable over the sample period, then fixed effects will tend to be inferior to pooled OLS or random effects methods, since time-invariant effects are differenced out by construction.

Table 2: Countries in the top-10 or bottom-10 of ranked Corruption Index values

Country	Corruption Rank	GDP Rank	Region
Sweden	1	9	Western Europe and North America
Norway	2	3	Western Europe and North America
New Zealand	3	21	Western Europe and North America
Denmark	4	10	Western Europe and North America
Netherlands	5	8	Western Europe and North America
Switzerland	6	5	Western Europe and North America
Singapore	7	2	South-East Asia
Canada	8	15	Western Europe and North America
Iceland	9	11	Western Europe and North America
Belgium	10	16	Western Europe and North America
Uzbekistan	131	95	Eastern Europe and post-Soviet Union
Egypt	132	77	North Africa & the Middle East
Madagascar	133	132	Sub-Saharan Africa
Afghanistan	134	125	South Asia
Tajikistan	135	114	Eastern Europe and post-Soviet Union
Guinea-Bissau	136	130	Sub-Saharan Africa
Turkmenistan	137	56	Eastern Europe and post-Soviet Union
Cameroon	138	111	Sub-Saharan Africa
Azerbaijan	139	54	Eastern Europe and post-Soviet Union
Chad	140	123	Sub-Saharan Africa

Missingness of Observations

An important facet of analyzing the sample is considering whether there are systematic patterns of missing data that could bias the results.

In many panel data settings, including this one, it is natural to expect that not all variables will be observed for every country in every year. If this data is missing at random, then missingness should not in general bias the results. But if data is missing in a way that is systemically related to the variables of interest, then our results could well be biased.

For instance, suppose that countries with lower values of corruption were more likely to have data reported (perhaps less corruption is associated with better recordkeeping and greater transparency). Alone, this correlation of corruption with missingness would present a distorted picture of summary statistics regarding corruption, but would not necessarily bias the regression results. Imagine though that the corrupt countries that had the least missing data were ones that were exceptionally wealthy despite their corruption (maybe they have large oil reserves, increasing funding for bureaucrats that track statistics, but also increasing the incentives for graft). In this case, higher rates of missing data for more corrupt countries with lower GDP will tend to bias the estimated effect of corruption on GDP upwards.

To investigate such potentially systematic patterns of missing data, I begin by looking at how many missing observations I have for key variables using the *codebook* command in Stata. From this command, I see that there are 128 missing observations for GDP and 44 missing observations for the political corruption index.

I then proceed to investigate systematic missingness in the data using regression analysis. In column (i) of Table 3, I perform OLS regression of an indicator for missingness (equal to one if either GDP or corruption measures are missing) on average values of GDP and corruption in each country over the period 1991-2016.[2] I would of course prefer to perform this regression using the true values of GDP and corruption, but by construction when the missing indicator is equal to one, that means at least one of the two variables is unobserved. By using the averages over the period of the data, I estimate whether countries that *tend to have* certain values of GDP and/or corruption are more or less likely to be missing.

From this OLS regression, I estimate that missingness is positively associated with both average corruption and average GDP. Specifically, a 1-centile change in corruption increases the probability that either GDP or corruption is missing by 0.07 percentage points according to the model.

In columns (ii) and (iii), I include country fixed effects to identify whether missingness is associated with GDP and corruption controlling for time invariant features of countries. By controlling for fixed effects, I identify whether within-country variation in the explanatory variable is associated with missingness. Controlling for fixed effects implies I cannot use the average values of GDP or corruption, so I instead run the regressions separate regressions controlling for GDP and corruption individually. I cannot control for GDP and corruption simultaneously because either GDP or corruption is missing by construction whenever the missingness indicator is equal to 1.

Table 3: Regression of missingness by variable

	(i) Missing GDP or corruption data (=1)	(ii) Missing corruption data (=1)	(iii) Missing GDP data (=1)
Average corruption index in country, 1991-2016	0.07***		
	(0.02)		
Average GDP per capita in country, 1991-2016	0.01**		
	(0.00)		
Log of GDP (2011 Constant Dollars PPP)		.	
Political corruption index			0.04
			(0.03)
Constant	-0.11**		0.04**
	(0.05)		(0.02)
Observations	3900	3772	3856

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Using this model, I find that a 1-centile increase in corruption is estimated to increase the probability that data (here GDP) is missing by 0.04 percentage points. On the other hand, whenever GDP per capita is non-missing, I find that I am not able to run the missingness regression. By looking at the cross-tabulation of corruption and GDP missingness in Table 4, I find that whenever corruption data is

[2] An OLS regression with a binary indicator as the dependent variable is called a linear probability model. In this framework, the coefficient on the explanatory variable is interpreted as the change in the probability of the dependent variable given a 1-unit change in the explanatory variable.

missing, GDP data is also missing. Hence, it is not possible to discern any association of "within" changes in GDP are associated with missingness of data.

Table 4: Cross-Tabulation Frequency Table of Missing GDP and Corruption Data

Observations missing	GDP data missing	GDP data non-missing	Total
Corruption data missing	3890	4345	8235
Corruption data non-missing	0	2565	2565
Total	3890	6910	10800

Model and Results

The empirical strategy for this analysis consists of regressing GDP on the Political Corruption Index corruption measure and controls, for each of pooled OLS, fixed-effects and random-effects regression. The analysis is repeated three times: without controls, with a set of controls capturing the long-run determinants of institutions, and finally with a set of time-varying production input controls.

Pooled OLS

In the basic Pooled OLS regression, I first regress the log of GDP per capita on the corruption index score without any controls. The regression is specified as follows:

Equation 1: Pooled OLS without controls

$$\log(GDP)_{i,yr} = \alpha + \beta_1 Corruption_{i,yr} + u_{i,yr}$$

Column (i) of Table 2 reports results for Pooled OLS regression without controls (using heteroskedasticity-robust standard errors). For the simple linear regression estimates, a 1-centile increase in the Political Corruption Index value is estimated to decrease GDP per capita by 2.77%, a very strong effect. This parameter estimate has a standard error of 0.04, so that the 95% confidence interval for the effect of corruption is 2.79% ± 2 x (0.04)%. Hence, under the model assumptions, I would expect the true coefficient to fall within the estimated confidence interval 95% of the time in repeated samples. That is, the estimate is significant at the $\alpha = 0.05$ level (and indeed, at the $\alpha = 0.01$ level).

The key assumption for unbiased inference in pooled OLS is that the error term (which includes the unobserved determinants of GDP per capita) is uncorrelated with the explanatory variables.[3] In the context of the simple linear regression we consider here, this means that political corruption has to be uncorrelated with all other determinants of GDP per capita. Researchers generally think of political corruption as an outcome of a complex social process, consequently it seems far-fetched to assume that corruption is neither caused by—or even simply correlated with—anything that also affects GDP.

For instance, closer cultural ties to the epicenters of Enlightenment Era philosophy during the 18th century might favor the development of better, non-corrupt government. Of course, closer cultural ties to the Enlightenment is true most of all for other European countries. Yet European countries also differ from other countries in their environment, history of industrialization and wealth, social norms and a host of other characteristics, which independently affect GDP per capita.

[3] For the estimator to also be the minimum variance unbiased estimator, the assumption of no heteroskasticity or autocorrelation would also have to be imposed. The assumption of homoscedasticity is highly unlikely in empirical settings like this one, however. As a result, I have used

While it is impossible to control for all potentially omitted variables, I can get some idea of the extent of the problem by controlling for the set of determinants of institutional quality discussed above. Specifically, I control for geographical/climatological determinants of institutions for former European colonies, as hypothesized by Acemoglu et al (2001). Acemoglu et al claim that more exploitative governments were created in areas with more hostile climates and note that the long-term impacts of these governmental origins can still be felt today. Thus, I take into account possible spurious correlation of corruption and GDP arising from the persistent correlation of corruption with the climate and disease environment of countries (two factors that seem likely to affect GDP). I also take into account the fact that when different colonizers settled different regions, the colonizers might have generated distinct cultural impacts in addition to the impacts on institutions / quality of government.

To do incorporate these considerations, I specify the regression with institutional determinant controls as:

Equation 2: Pooled OLS with institutional determinant controls

$$\log(GDP)_{i,yr} = \alpha + \beta_1 Corruption_{i,yr} + \beta_2 BritishCol_i + \beta_3 FrenchCol_i + \beta_4 Latitude_i + \beta_5 Temp_i + \beta_6 Malaria_i + \beta_7 YellowFever_i + u_{i,yr}$$

Table 5 reports the estimates of this model. The estimated effect of political corruption on GDP per capita is smaller than in the no controls case, at 1.24%. The fact that controlling for the determinants of institutional quality more than halves the estimated effect if corruption is a clear indication that the simple regression suffered from omitted variable bias.

Turning to the control variables, I estimate a one-degree increase in latitude to decrease GDP per capita by 2%, while an increase in the mean temperature by 1 degree is estimate to decrease GDP per capita by 7%. The presence of yellow fever is estimated to increase GDP per capita by 25%.[4] A 1-centile increase in the Malaria index is estimated to reduce GDP per capita by 1.63%. Having previously been a British colony is estimated to increase GDP by 27%. Finally, having been a former French colony is estimated to increase GDP per capita by about 12%. For each of these controls, when using heteroskedasticity-robust standard errors, the estimates are significant at the $\alpha = 0.01$ level, except for the former French colony indicator, which is significant at the $\alpha = 0.05$ level.

The relevant assumption of the pooled OLS equation with institutional determinant controls is that corruption is uncorrelated with unobserved determinants of GDP *conditional on* the institutional determinant controls. Of course, I have already mentioned that there are likely other institutional determinants that I haven't controlled for here – like proximity / shared cultural legacy to Enlightment philosophy or Communist rule. There are likely many other such omitted long-run determinants of GDP correlated with corruption. Generally, these correlates of long-run quality of government are likely to not change over the duration of the panel.

[4] Note that the point estimate is 0.22, hence it is tempting to interpret the change as 22%. But remember that log-linear models imply that the *marginal effects* are interpreted as (100 x β)% changes in the dependent variables. These marginal effects are good approximations when the implied changes in the dependent variable are small (providing a very close approximation for implied changes up to 5% and a reasonably close approximation for values up to 20%). For larger variables, the implied change in the dependent variable is $e^{\Delta X * \beta} - 1$. Hence, the estimated effect for a 1-unit change in the Yellow fever indicator (ie going from no Yellow fever to yellow fever present) is $e^{1*0.22} - 1 \approx 1.25 - 1 = .25$, or 25%.

In addition, however, there are many omitted variables affecting GDP that are likely to change over time. Some examples of this include changes to the economy (eg increasing inequality, industrial shocks or transformation), dynamics of social cohesion (e.g. civil strife), changing demographics (eg a more educated population), or changing social values. Many of the time-varying omitted variables I just mentioned are potentially problematic to include, however, because while some of these developments may occur independently of corruption, they could vary in part as a result of corruption. Because of this concern, I have chosen to use time-changing variables (relating to the economy) which are more plausibly exogenous to changes in corruption. First, I use lagged population (here, the population five years ago) as a time-varying measure of labor inputs into the economy. By using the 5-year lag of population, this measure should not be affected by contemporary variation in corruption. Additionally, I use the value of oil and natural gas endowments (as measured by production). The value of natural resource production inherently varies over time as prices of these inputs fluctuate, but arguably, corruption should not affect these production values. Despite not being outcomes of corruption, natural resources are still likely correlated with corruption due in part to the Acemoglu et al hypothesis about colonization. Moreover, one can further imagine that resource endowments are potential determinants of corruption because the greater the natural wealth of a country, the greater the incentives to profit from these resources via corruption.

For the sake of comparison, I estimate the regression using production input factor controls while omitting the previously selected institutional determinant controls. If I were striving to reduce omitted variable bias, I would likely want to include both – but using different sets of controls is instructive for comparing results between models.

Equation 3: Pooled OLS with production input factor controls

$$\log(GDP)_{i,yr} = \alpha + \beta_1 Corruption_{i,yr} + \beta_2 Oil_{i,yr} + \beta_3 Gas_{i,yr} + \beta_4 Pop_{i,yr} + u_{i,yr}$$

Column (iii) of Table 5 presents the estimates for pooled OLS regression with natural resource value controls. In this specification, I find a similar effect of corruption as in the no-control case, at 2.8%. I moreover find that a 100 million person increase in population is estimated to increase GDP per capita by approximately 15%, while an increase in oil production by 10 billion US (2011) dollars is estimated to increase GDP per capita by 9%. All of these estimated effects are significant at the $\alpha = 0.01$ level. Gas production is not estimated to affect GDP per capita.

The key assumption underpinning these results is that corruption is uncorrelated with unobserved determinants of GDP after controlling for the time-varying production input factor controls. Of course, we have good reason to doubt this given that we have not controlled for the previously identified long-run correlates of corruption. But even had I used both sets of controls, each are just examples of relevant controls; they are by no means exhaustive. There are an abundance of other time-varying and long-run determinants of GDP that seem likely to correlate with corruption.

Pooled OLS with Clustered-Standard Errors

Besides the concern with bias in the Pooled OLS regression, there is also the issue of whether appropriate assumptions about the error structure have been made to allow for correct inference. If intragroup correlation is important, then my use of heteroskedastic-robust standard errors will lead to incorrect conclusions during hypothesis testing—because the true standard errors may be much larger (and thus t and F-test statistics much smaller).

On the basis of theory, we likely expect intragroup correlation in the error term to be quite important, as there are many unobserved characteristics about country that are likely to correlate highly from one period to the next while having an impact on GDP (regardless of any possible correlation with corruption). When this intragroup correlation is high, it is generally necessary to account for this serial correlation in the error term by using clustered standard errors.

The estimated standard errors when using clustered standard errors (at the country level) are reported in brackets in Table 5 (heteroskedastic-robust standard errors are in parentheses). The standard errors produced using cluster-robust methods are generally about five times higher than the heteroskedastic robust estimates, suggesting that inference is very mistaken when failing to account for intragroup correlation. Indeed, statistically significant estimates for latitude, presence of yellow fever, British/French colonial origin, and population all become non-significant when using cluster-robust standard errors.

Fixed Effects Regression

In panel settings, instead of trying to proactively observe and control for every type of omitted variable bias, one can control for any time-invariant omitted variables of a country by allowing for country fixed effects.

The fixed effects model without controls is specified as:

Equation 4: Fixed effects regression without controls

$$\log(GDP)_{i,yr} = \delta_i + \tau_{yr} + \beta_1 Corruption_{i,yr} + u_{i,yr}$$

Where δ_i denotes the country-specific fixed effect and τ_{yr} denotes the year fixed effect.

Column (i) of Table 6 reports results for the fixed effects regression using only corruption in addition to the country and year fixed effects. Using this model, I obtain much smaller estimates of the effect of corruption on quality of government than in the results from pooled OLS. Specifically, my results for this specification indicate that a 1-centile increase in corruption is estimated to decrease GDP per capita by only 0.28%, about an order of magnitude smaller than in the pooled OLS without control variables. The effect of corruption is still statistically significant, but only at the $\alpha = 0.1$ level.

The key assumption of the fixed effects regression is strict exogeneity, which states that the error term is uncorrelated not only with explanatory variables in the contemporary period, but also in all other time periods, conditional on the fixed effects.[5] Two major implications arise from strict exogeneity.

First, using country fixed effects controls for any time-invariant omitted variable bias specific to each country. This means that I need only worry about omitted variables that vary over time. When also including year fixed effects as I have, then any omitted variable that affect all countries the same in a given period of time are also controlled for. In this case, I can be more specific in stating that the only omitted variable in the fixed effects regression are omitted variables that vary *differently between countries* over time.

The second major implication of strict exogeneity is that correct timing of the variables is important. If, for example, an explanatory variable affects the outcome with a lag not accounted for in the regression, then the error term in a given period will correlate with the explanatory variable in a previous period, violating strict exogeneity. In the analysis here, I have omitted use of lags. Largely, this is because

[5] Wooldridge states this requirement as zero mean of the error term conditional on the explanatory variables (including the fixed effects) for all periods.

doing so makes the assignment more complex – and it is quite simply hard to figure out which lags should be used – one would certainly imagine that corruption has a persistent effect into future period, as might production factor inputs to a lesser degree. As a result, I am very confident that the regression specifications applied here violate strict exogeneity.

In addition to concern about timing, however, there is also basis for concern about strict exogeneity on the basis of hypothesized omitted variable bias from time-varying sources. Once again, the various factors of economic upheaval, changes in social cohesion, demographics, and social mores all present themselves as possible omitted variables that vary across time differentially between countries. I will preliminarily investigate a small sampling of these confounders using the input factor controls already identified.

With the population and natural resource production input factor controls, the fixed effects regression appears as follows:

Equation 4: Fixed effects regression with input factor controls

$$\log(GDP)_{i,yr} = \delta_i + \tau_{yr} + \beta_1 Corruption_{i,yr} + \beta_2 Oil_{i,yr} + \beta_3 Gas_{i,yr} + \beta_4 Pop_{i,yr} + u_{i,yr}$$

Column (ii) of Table 6 presents the estimated results for the fixed effects regression with input factor controls. The estimated effect of corruption on GDP per capita is very similar to the fixed effects specification without controls, with a 1-centile change in corruption estimated to reduce GDP per capita by 0.25% (also significant only at the $\alpha = 0.1$ level). The estimated effects for control variables is also quite different between Pooled OLS and fixed effects regression. Whereas in Pooled OLS, the estimates for population and oil production were statistically significant (but not gas production), in the fixed effects regression it is only the estimated effect for gas production that is statistically significant (at the $\alpha = 0.1$ level). The magnitude of the impact is generally small, however, with a ten billion dollars in natural gas production estimated to increase GDP per capita by only 0.59%.

Three further observations are worth noting about analysis using the fixed effects model. First, I have not included estimates for the regression specification with institutional determinant controls. Because these institutional determinants are time-invariant (based on a single set of observations, generally in the past), then the effects identified by these variables are already controlled for by the country fixed effects. Indeed, when trying to including these regressors in Stata, they are dropped from the regression because they are collinear with the fixed effects terms.

Second, the standard error of the estimates in the fixed effects regression are much larger. This is reasonable, given that pooled OLS uses all of the within and between variation, while the fixed effects estimator uses only the within variation. In part, this may explain the weaker statistical significance of results in the fixed effects model, although weaker significance is also due to smaller point estimates in this case.

Finally, it is important to note that fixed effects regression might suffer from worse omitted variable bias than pooled OLS, despite controlling for significantly more omitted variables (all time-invariant omitted variables for each country and country-invariant omitted variables for each year). The reason for this is that we are ultimately concerned with the aggregate correlation between the explanatory variable and unobserved determinants of GDP). As already mentioned, researchers think that long-term or time-invariant characteristics determine much of the variation in quality of government (as verified using the within-and-between variance decomposition). Fixed effects regression isolates only within-country variation in corruption that occurs across years of the sample. That is to say, the variation used

in this analysis is only short-term changes in corruption for each country, which is unexplained by either long-term determinants of corruption or common trends in corruption across country. Why is corruption changing in this case? In particular, does it seem likely that this idiosyncratic variation in corruption over time is more or less related to the error term?

In pooled OLS, omitted variable bias is based off the correlation between corruption (inclusive of long-run determinants) and an error term that includes both time-invariant and time-varying determinants of GDP. In fixed effects regression (with year fixed effects), we are only comparing idiosyncratic variation in corruption for each country over time against omitted variables that also vary differently between countries over time. It is very possible that the strength of the correlation could be greater for fixed effects. For example, I may suspect that short-term increases (/reductions) in corruption reflect a deteriorating (/improving) political environment due entirely to shocks to either the social or economic conditions of the country. In this case, the within-variation in corruption used in fixed effects is merely an outcome of the omitted variable. Hence, even though I have controlled for more types of omitted variables, I have made the correlation between the explanatory variable and the error worse.

Random Effects Regression

Finally, I estimate the effect of corruption on GDP per capita using the random effects model. The random effects model is used to produce more efficient estimates than pooled OLS (by correcting for serial correlation in the error term) when the researcher assumes that the individual effect is uncorrelated with the explanatory variables.

The random effects model with no controls is specified as:

Equation 5: Random effects regression without controls

$$\widetilde{\log(GDP)}_{i,yr} = \check{\tau}_{yr} + \beta_1 \widetilde{Corruption}_{i,yr} + \check{u}_{i,yr}$$

Where τ_{yr} denotes the year fixed effect and for any variable x_{it}, $\widetilde{x_{it}} = x_{it} - \theta\,\bar{x}_i$, with θ measuring the strength of the within-group correlation in the error term.

Table 7, column (i) reports results for the random effects regression without controls. The point estimates for this regression are closer to fixed effects regression than random effects, with 1-centile increase in political corruption index estimated to reduced GDP per capita by 0.38%. Computationally, the random effects is an intermediate case between pooled OLS and fixed effects regression, depending on the strength of the intragroup correlation (θ). If $\theta = 0$, the random effects estimator is identical to the pooled OLS estimator, while in the case of $\theta = 1$, the random effects estimator is identifical to the fixed effects estimator. In the random effects regression of GDP on corruption without controls, the median estimate of $\hat{\theta}$ is 0.96,[6] hence it is not surprising that the random effects estimator produces results that are very similar to the fixed effects estimator.

The key requirement of random effects regression is that one need not control for the unobserved time-invariant country effects, because they are assumed uncorrelated with corruption (thus omitting them will not cause omitted variable bias). Under this assumption, random effects allows the researcher to correct regression estimates so that the residuals become homoscedastic. Hence, while pooled OLS only adjust inference to take into account the fact that the error term displays serial correlation, random effects essentially tries to remove the serial correlation.

[6] Median estimates for institutional determinant sets and factor input control sets are 0.93 and 0.95,a respectively.

However, there is very good reason to believe that time-invariant omitted determinants of GDP are correlated with corruption (as mentioned with the long-run determinants of institutions). Hence, once again, the key assumption for unbiasedness or consistency is unlikely to hold.

What is more unclear is whether random effects is likely to be more or less biased than pooled OLS or fixed effects. Since random effects is an intermediate case between pooled OLS and fixed effects regression, OLS seems likely to be less biased if the correlation with the error term is weaker using between-variation, while fixed effects regression is likely less biased if correlation with the error term is weaker when using within-variation, and random effects should generally fall somewhere inbetween. Since the median estimate of $\hat{\theta} = 0.96$ for the no control variables case, the random effects estimator will generally have a bias that is similar to fixed effects.

If the random effects was consistent, this implies all three estimators would be consistent while random effects would be the most efficient of the three. But once again, random effects is almost certainly not consistent given expected correlation between corruption and unobserved time-invariant determinants of GDP.

To control for further sources of omitted variable bias, I also run random effects regression separately using institutional determinant controls and factor input controls, as in the pooled OLS and fixed effects regressions.

The random effects regression with institutional determinant controls is specified as follows:

Equation 5: Random effects regression with institutional determinant controls

$$\widetilde{\log(GDP)}_{i,yr} = \check{\tau}_{yr} + \beta_1 \widetilde{Corruption}_{i,yr} + \beta_2 \widetilde{BritishCol}_i + \beta_3 \widetilde{FrenchCol}_i + \beta_4 \widetilde{Latitude}_i + \beta_5 \widetilde{Temp}_i + \beta_6 \widetilde{Malaria}_i + \beta_7 \widetilde{YellowFever}_i + \check{u}_{i,yr}$$

Where τ_{yr} denotes the year fixed effect and for any variable x_{it}, $\widetilde{x_{it}} = x_{it} - \theta\, \overline{x_i}$, with θ measuring the strength of the within-group correlation in the error term.

As in the case of pooled OLS, this specification will control for more omitted variable bias stemming from unobserved correlates of long-run institutional determinants of corruption, i.e. characteristics that are more stable over time. The estimated effect of political corruption on GDP per capita is largest among the fixed and random effects regressions using this specification, with a 1-centile increase in corruption reducing GDP per capita by 0.55%. Statistical significance levels of the control variables in this case is the same as in the pooled OLS case with clustering (only mean temperature and Malaria index rates are significant, both at the $\alpha = 0.01$ level).

Finally, the random effects regression with factor input controls is specified as follows:

Equation 5: Random effects regression with institutional determinant controls

$$\widetilde{\log(GDP)}_{i,yr} = \check{\tau}_{yr} + \beta_1 \widetilde{Corruption}_{i,yr} + \beta_2 \widetilde{Population}_{i,yr-5} + \beta_3 \widetilde{Oil}_{i,yr} + \beta_4 \widetilde{Gas}_{i,yr} + \check{u}_{i,yr}$$

Where τ_{yr} denotes the year fixed effect and for any variable x_{it}, $\widetilde{x_{it}} = x_{it} - \theta\, \overline{x_i}$, with θ measuring the strength of the within-group correlation in the error term.

Factor input controls will control for more of the time-varying unobserved correlates of corruption. In this specification, the estimated effect of a 1-centile increase in the corruption index is a 0.41% decrease in GDP per capita. The controls for population and oil are not significant in this specification, while a $10 billion increase in gas production is estimated to increase GDP per capita by 1%.

Across all random effects regression, it is worth noting that the standard error estimates of the coefficients are much larger than pooled OLS (with heteroskedastic errors) but still smaller than fixed effects regression. This once again is an expected result, given that pooled OLS uses all of the variation, fixed effects regression uses only within variation, and in this case random effects uses within-variation and a small amount of the between variation.

Conclusion

The results offer modest support for a negative relationship between corruption and GDP growth. Pooled OLS estimates suggest a roughly 1.24-2.80% negative effect of 1-centile change in the corruption measure on GDP, while random and fixed-effects estimates are smaller at about 0.28-0.55% depending on the controls.

An important question underlining all these approaches, however, is whether the results should be trusted? In short, no. Without resorting to instrumental variables, there is no strong basis for assuming any quality of government variable is exogenous. Indeed, there is ample reason to suspect that a variety of factors directly affecting GDP also influence quality of government. A pooled OLS design rests entirely on the assumption of exogeneity; hence, the results are most likely biased. Random effects suffers from largely the same shortcomings, given that the we are required to assume that unobserved country specific effects are uncorrelated with the quality of government measure.

Often, fixed effects regression will provide the most convincing case for unbiasedness, because it allows the researcher to control for any time-invariant variables that may bias the results. But while the fixed effects regression controls for time-invariant omitted variables, many omitted variables that seem relevant vary across both time and countries, including the possibility of reverse-causality between GDP and quality of government.

The concern about time-varying omitted variables is especially important when one considers that we have controlled for time-invariant effects of corruption by using fixed effects. Hence, the variation in corruption observed here is presumed unrelated to long-run determinants of quality of government that are typically considered most important. What is causing these short-run increases or decreases in corruption? It is unclear, but seems susceptible to a host of potential confounders.

Consequently, each of the three methods have shortcomings that are good reason for doubting the relevant exogeneity assumptions and therefore any resulting causal inference. In the next assignment, we will investigate instrumental variables design, which attempts to address these exogeneity concerns more directly.

Table 5: Pooled OLS Regression of GDP and corruption

	(i) No controls	(ii) Institutional determinant controls	(iii) Factor input controls
Political corruption index	-2.77*** (0.04) [0.19]	-1.24*** (0.06) [0.30]	-2.80*** (0.04) [0.17]
Latitude of capital (absolute value)		-0.02*** (0.00) [0.01]	
Mean temperature		-0.07*** (0.00) [0.02]	
Yellow fever present today		0.22*** (0.03) [0.15]	
Malaria index in 1994		-1.63*** (0.05) [0.23]	
Former British colony		0.24*** (0.04) [0.19]	
Former French colony		0.12** (0.05) [0.22]	
Population (Lagged, 100 millions)			0.15*** (0.04) [0.14]
National oil production (in 2014 dollars, 10 billions)			0.09*** (0.01) [0.02]
National gas production (in 2014 dollars, 10 billions)			-0.00 (0.01) [0.03]
Constant	10.32*** (0.02) [0.10]	11.47*** (0.13) [0.67]	10.22*** (0.02) [0.10]
Observations	3772	1423	3267

Heteroskedasticity-Robust standard errors in parentheses
Cluster-robust standard errors in brackets.
P-values stars reported based on heteroskedasticity-robust standard errors.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Fixed Effects Regression of GDP and corruption

	(i) Log GDP – No controls	(ii) Log GDP – Natural Resource Value Controls
Political corruption index	-0.28*	-0.25*
	(.17)	(.14)
Latitude of capital (absolute value)		
Mean temperature		
Yellow fever present today		
Malaria index in 1994		
Former British colony		
Former French colony		
Population (Lagged, 100 millions)		-.15
		(.2)
National oil production (in 2014 dollars, 10 billions)		.000016
		(.0048)
National gas production (in 2014 dollars, 10 billions)		.0059**
		(.0027)
Constant	8.8***	8.9***
	(.089)	(.081)
Observations	3772	3160

Cluster-robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Random Effects Regression of GDP and corruption (no/domestic controls)

	(i) No controls	(ii) Institutional determinant controls	(iii) Factor input controls
Political corruption index	-0.38**	-0.55***	-0.41***
	(0.16)	(0.16)	(0.13)
Latitude of capital (absolute value)		-0.02	
		(0.01)	
Mean temperature		-0.08***	
		(0.02)	
Yellow fever present today		0.25	
		(0.16)	
Malaria index in 1994		-1.78***	
		(0.25)	
Former British colony		0.33	
		(0.21)	
Former French colony		0.13	
		(0.22)	
Population (Lagged, 100 millions)			-0.07
			(0.15)
National oil production (in 2014 dollars, 10 billions)			0.00
			(0.00)
National gas production (in 2014 dollars, 10 billions)			0.01**
			(0.00)
Constant	0.00	11.38***	0.00
	(.)	(0.70)	(.)
Observations	3772	1370	3160

Cluster-robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$