**CSCI 3346, Data Mining**

**Spring 2015**

**Term Project Written Report**

**Airen Surzyn**

**Abstract**

This project explores one of many theories behind the notorious resource curse, a pressing problem in development economics and international relations. Specifically, I tried to find a plausible positive correlation between institutional quality and economic growth, a theoretical connection posited by Halvor Mehlum, Karl Moene, and Ragnar Torvik(2006). However, attempting to find objective proof to social science problems is fraught with complications. The curse of dimensionality, known to all areas of data science, becomes exaggerated when considering the impossibility of identifying all factors of variability in such a problem. My main concerns for the project were derived from the necessary approximations of such things as ‘institutional quality’ that would likely blunt my instruments of analysis. These concerns proved valid as the linear models I created hardly suggest plausible correlation between the two considered variables. Should the characterizations if ‘institutional quality’ and ‘economic success’ been known to be accurate, then I might have been able to dismiss the correlation given my results. However, given the poor fit of my models, I am left questioning the accuracy of my metrics before the quality of my methods. The improvements observed between my first and final model are indicative of the power of the intermediary steps of data mining in coaxing meaning from otherwise incoherent datasets. Though I did not prove a correlation as I had originally intended, I was able to make strides in identifying important variables for consideration in questions surrounding the resource curse.

**Introduction**

One of the major observations in late 20th century international relations theory has been the tendency of resource abundant developing countries to struggle with economic development despite their natural endowments. The connection between resource abundance and stagnated economic growth was first demonstrated by Jeffrey Sachs and Andrew Warner(1995) and has since become a focus of development economics. Throughout the 20th century countries such as Mexico, Venezuela, and Sudan have failed to fully capitalize on their immense natural advantage while others, with little to no resources such as Singapore and Taiwan have done far better with far less.

In this paper I look at a theory proposed by Halvor Mehlum, Karl Moene, and Ragnar Torvik(2006) which suggests that institutional quality is positively correlated with economic growth. In other words, countries that provide more efficient governance and adjudication are more likely to experience sustained economic growth. Mehlum, Moene, and Torvik cite Botswana as a prime example of a country abundant in resources and with a high-functioning bureaucracy that has managed to maintain one of the highest growth rates over the last half century[[1]](#footnote-2). Inefficient or ineffective institutions are named the bane of sustained development as they open opportunities for corruption and the improper allocation of revenue from natural resource harvesting.

The theory posited by Mehlum, Moene, and Torvik is a question of establishing plausible correlation between the two variables of institutional quality and economic growth. Measuring economic growth is a fairly standardized, though imperfect, metric most often gauged in terms of GDP. Of the two variables, I have more confidence that this variable will provide a better approximation of economic realities. Measuring institutional quality is, on the other hand, a far more difficult task that calls into question the applicability of objective analysis techniques in such questions. The value of Positivist approaches in International Relations and the broader social sciences is a consistent point of controversy for exactly this reason. As has often been mentioned, the curse of dimensionality poses serious difficulties to any study. This problem is exaggerated in the social sciences where variables are innumerable and often unidentifiable. Evaluations of ‘good’ institutions constitute a field of study in themselves. For the most appropriate measurement of this variable I relied on the domain expertise of the World Bank and its affiliates. Of data on institutional quality that I was able to find, the Worldwide Governance Indicators provide the most consistent and standardized measurements of institutional quality. There are six individual measurements: Voice and Accountability, Political Stability, Control of Corruption, Regulatory Quality, Rule of Law, and Government Effectiveness. In my study, I aggregated these into one composite metric that would be representative of ‘institutional quality.’

My variable for economic growth would be the average percent change in GDP per annum between the years 1996 and 2013[[2]](#footnote-3). Institutional quality would be measured by my composite WGI variable. I picked my ‘resource-dependent’ countries based on the same guidelines as Mehlum, Moene, and Torvik. In their study, ‘resource-dependent’ referred to countries whose GDP was comprised of more than 10% resource exports. Considering different time periods, their study included 42 countries. I selected my countries based on the year 1996(the first year that the World Governance Indicators cover) which meant my country dataset would be composed of 54 countries.

**Methods and Results**

The data I would be using was easy to obtain and required minimal curating. The World Bank’s Databank provided everything from the WGI statistics to measurements of GDP. My pre-processing and data analysis were completed jointly in MATLAB and Microsoft Excel.

The first task was to determine my country dataset. I scanned the ‘Total natural resources rents (% of GDP)’ of the year 1996 and extracted all countries whose percentage was higher than 10%(the boundary of resource dependence). Having the countries to be considered, I retrieved the information for the other metrics. My timescale was 1996 to 2013, and for this range I retrieved percent change in GDP per annum and the six World Governance Indicators for each of the 54 countries. Occasional missing values in percent change in GDP were substituted with the average of the other years of the country’s percent change in GDP. Next, I dealt with combining the World Governance Indicators to form my ‘institutional quality’ composite metric. There seemed no reasonable way to assume emphasis on any of the individual measurements, so weighting them in the composite stat seemed unwise. Thus, I averaged the values and assigned each singular statistic to its respective country. At this point, I had my country dataset, average percent change in GDP, and my institutional quality metric assigned to appropriate vectors in MATLAB.

I began the study in similar fashion to Mehlum, Moene, and Torvik with a linear regression model. The question being one of correlation, looking for a linear relation among the data points seemed to be the most sensible method. Correlation would be reasonably established if it would be possible to fit the data points to a common line with minimal variance. Mehlum, Moene, and Torvik’s assertion of the existence of the relation between institutional quality and economic growth was based on a similar linear regression, though with different metrics. I used MATLAB’s *LinearModel.fit* method to plot a linear regression using the WGI metric as the first input and the percent change in GDP as the second. This yielded the regression model seen in Figure 1. This first model was poorly fit, not inconsistent, however. with that of Mehlum, Moene, and Torvik. As can be seen in Figure 1 and Figure 2, the plotted data points resemble noise more than any discernible pattern. Examining the objective details of my model, it yielded an R square of .0161 and a p-value of .36(See Appendix for further detail). So, as of yet, the null hypothesis remained and a plausible correlation was far from established. Thus, I set about trying to massage some sort of meaning from the dataset and model I had created.

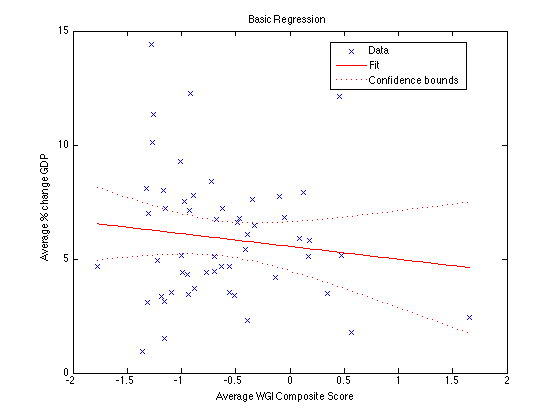


Fig. 1

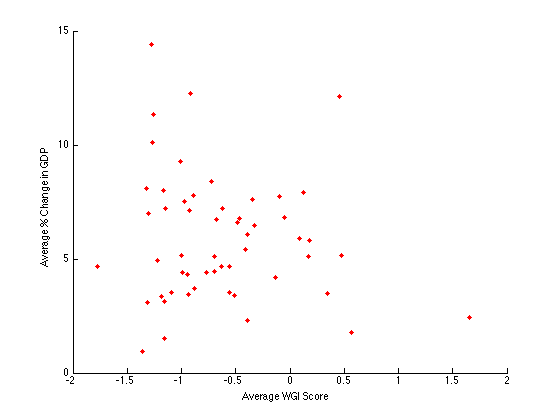


Fig. 2

Examining the naked scatter plot of Figure 2, I searched for any signs of clustering. As can be seen on the plot, there were no discernible groupings other than a slight density of points between -1.5 and 0 on the x-axis and 2.5 and 8 on the y-axis. Thus, I considered methods of feature creation so as to attempt to separate the points using another dimension of consideration. I arrived at my feature by examining the most extreme points on the graph, notably the point on the far right and bottom of the plot representing Norway. Norway, in reality a very economically successful and institutionally stable country was represented on my graph as an institutionally sound, though economically unsuccessful country based on its low percent GDP growth. However, this was a completely normal phenomenon of developed countries. GDP growth rates of 10% are largely unsustainable for developed economies and tend to hover around two or three percent for a successful economy. Noting this, I decided to create new class labels for the data points that would separate them based on their overall economic levels of development. Based on the World Bank’s classification of national income from Low to High, I assigned discrete class groupings to each country. Applying this new feature to my data points I performed a *gscatter* of the same dataset to produce a plot with slightly more discernible features(seen below).

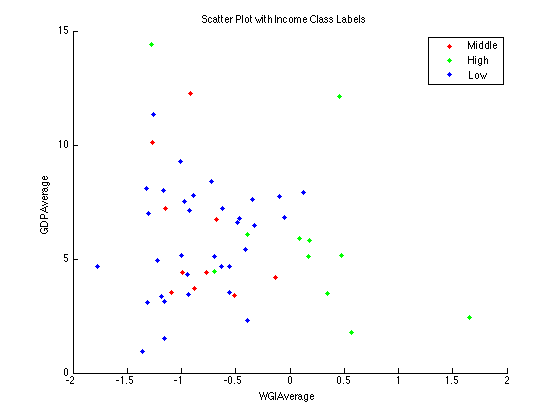


Fig. 3

Countries of the High Income class label tended to have higher institutional quality and lower average percent change in GDP. Again, the low GDP growth in High Income countries is a result of a more developed, less volatile economy and can hardly be taken as an indication that greater institutional quality leads to lower rates economic growth. However, this seemed to be a misleading feature of my plot which would need to be addressed. The Low Income and Middle Income countries were nearly inseparable existing in the same space with seemingly little separation.

Reconsidering my dataset, I looked at the data points in the top left space of my plot(low institutional quality, extremely high growth rates) and arrived at another important conclusion regarding the representation of the dataset. After some background research into the specific cases themselves, I found that many countries such as Angola or Equatorial Guinea had overwhelmingly large dependencies on certain resource sectors of their economies. The economies of these two countries are largely dependent on crude oil exports and being a very lucrative commodity have profited in disproportionate amounts from this sector vis a vis the rest of their economy. This illuminated a detail I had overlooked which was the degree to which resource dependence varied among the countries in my dataset. Already having the requisite data to make the adjustments I revived the aforementioned ‘Natural Resource Rents(% of GDP)’ statistic and decided to use it as a weight for the data points. Through simple multiplication I applied the resource rent percentage to the percent change in GDP. Natural groupings in the dataset, though still imperfect, began to appear. As seen below, the formerly interspersed Low and Middle Income data points achieved some separation with the weighting of the data points.

Given the small number of data points for Middle and High Income countries, I felt any further regression or consideration of these particular groupings would be spurious and fruitless. I tried applying linear models to each, but it resulted in lines with very large slopes that would clearly not be representative of reality. However, the feature creation of income-based class labels and the weighting of the data points had given me an adequately large and much more specified dataset for Low Income countries. Further, the resource curse is largely a problem of development economics. Thus, I had not compromised the original goals of the study by pursuing further inquiry within a subset of the original dataset.

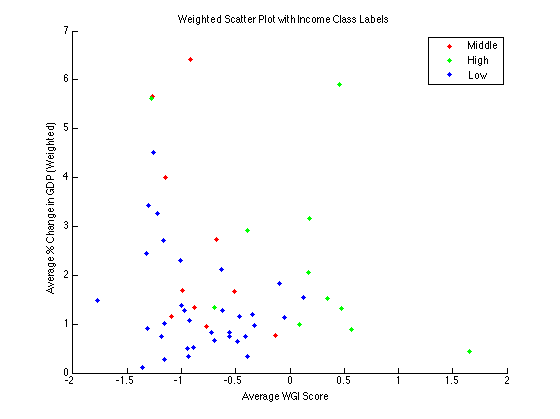


Fig. 4

With my new dataset ready, I reapplied the linear regression strategy to the new subset of Low Income, weighted data points(Fig. 5). As one can see below, it was still not a fit from which one could draw any conclusion regarding the correlation between my two variables. However, it was notably better than my initial regression attempt with an R-squared value of .0719 and a p-value of .131(see Appendix for further detail). The negative slope of the graph in fact implied the opposite, suggesting a negative correlation between institutional quality and economic growth. However, the fit of the model was not strong enough to provide plausible support for what would be still a dubious conclusion.

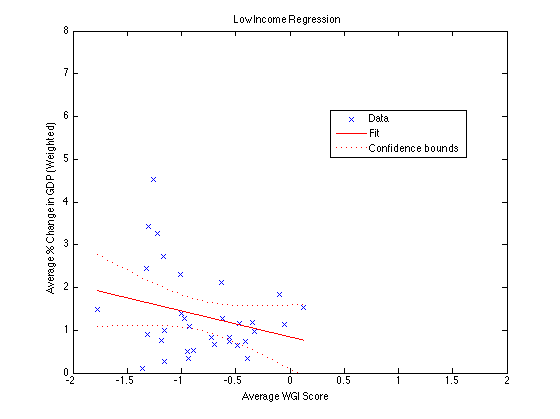


Fig. 5

**Conclusion**

My expectations were reasonably low coming into the project. That is to say, I certainly did not expect to solve the resource curse or provide scholarly level proof of newfound truths in development economics. Rather, I had hoped to expose myself to the process and the considerations that would come with trying to approach the social sciences with Data Mining techniques and practices. And in this regard I was successful. My initial model was an obviously poor fit, but very much committed to the scope of my task, I persisted in pursuing meaning in what seemed to be a largely incoherent dataset. Examining the data instances through data visualization techniques, I was able to identify certain characteristics of the dataset that lead me to pursue feature creation and to slightly alter the measurements I was using in my plotting. This, in turn, led me to identify a smaller and more useful subset of data within my original, broader consideration. Though the linear model I fit to this dataset was still to weak to establish a believable correlation in either direction, it had improved greatly relative to the other dataset. However, this exercise has led me to the believably salient identification of income levels as an inherently important characteristic in the consideration of such questions as the resource curse. By isolating data points with this common feature I was able to achieve a markedly better fit for my model which indicates its usefulness. More than anything, this project highlighted the importance of working within the data to identify key characteristics and information upon which one can turn to look for important information that may be initially hidden. Going forward, I will need to work on refining the methods with which I approach the evaluation of qualitative metrics of things such as ‘institutional quality’. Refining these approximations leads the way to more successful endeavors of data mining in the social sciences.

**References**

<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>; Regression Analysis: How Do I Interpret R-squared and Assess the Goodness of Fit?; Frost, Jim; 2013

Mehlum, Halvor; Moene, Karl; Torvik, Ragnar; *Institutions and the Resource Curse*; The Economic Journal; Royal Economic Society; 2006

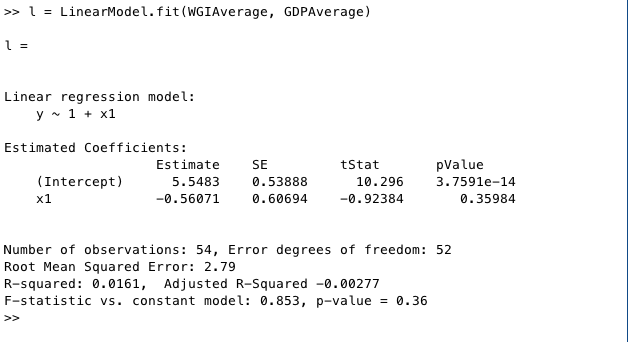
Rodrik, Dani; *The Future of Economic Convergence*; Harvard University; 2011

Sachs, Jeffrey; Warner, Andrew; *Natural Resource Abundance and Economic Growth*; Harvard University; November, 1997

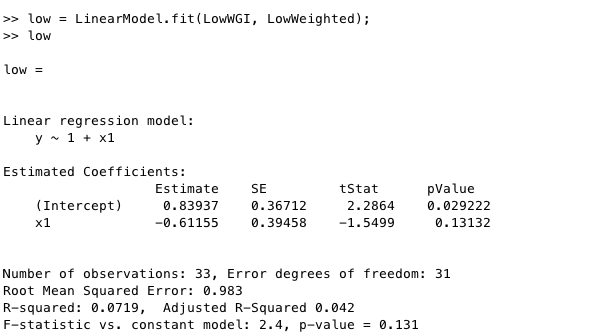
Sachs, Jeffrey; Warner, Andrew; *Natural Resources and Economic Development: The Curse of Natural Resources*; Harvard University; 2001

**Appendices**

First Regression Model:



Weighted Low Income Model:



1. Institutions and the Resource Curse, page 3 [↑](#footnote-ref-2)
2. The timespan of the WGI measurements [↑](#footnote-ref-3)