Zipf's Law and Its Correlation to the GDP of Nations

Rachel K. Skipper (Frostburg State University)
Mentor: Dr. Jonathan Rosenberg, Ruth M. Davis Professor
University of Maryland, College Park

Abbstract

This study looks at power laws, specifically Zipf's law and Pareto distributions, previously used to describe city size distribution, income distribution within firms, and word distribution within languages and documents among other things, and Gibrat's law describing growth rate. This study seeks to discover if Zipf's law can also be used to model the distribution of GDP's worldwide using Gibrat's law as a justification. The simplest method to determine Zipf's law's applicability, and the one used in this study, was to create a log log plot, plotting rank versus size of the GDPs. Using that plot, Zipf's law was verified through two criteria. First the plot must appear linear and second it must have a slope of -1. For the purpose of this study, the data looked at was for all countries and then countries split into categories of emerging economies and advanced economies for the years 2005, 2006, 2007, and 2008. The results of this study showed that all countries and countries with emerging economies did not appear linear on the log log plot while advanced economies appeared linear with a slope roughly -.70, suggesting that GDP distribution of advanced economies instead follow a Pareto distribution. Advanced economies also showed a significantly smaller variation in growth rates over the four years as implied by Gibrat's law. This was used as a possible explanation for the distribution discovered.

Introduction

Problem Statement

Today, countries interact constantly with one another sending products and services around the world. Despite these international dealings there still remains a disconnect between the wealthiest and the poorest countries. Some countries seem to have entirely successful and large economies while it seems as though other nations flounder and can barely sustain their populations. A few countries experience extreme wealth while others survive with extreme poverty. Although this may not be a problem for countries with large economies, smaller economies and limited resources pose a significant dilemma for others.

The implications of the massive variance between economies are numerous and go beyond the national level. Haddad, et. al. (2003) noted that malnutrition reduction rates within a country followed closely with both household income and national levels. Filmer and Pritchett (1999) indicated by their study that income per capita and inequality of income distribution, among other factors, explain ninety-five percent of cross-national variations in mortality. Others have found that foreign trade, investment, and debt dependency, all interconnected factors in GDP, negatively affect infant mortality rates (Shen & Williamson, 2004). With this in mind, the problem thus emerges to explain and possibly try to mitigate the differences between the very wealthy countries and the much poorer countries.

Purpose of Study and Research Questions

With the significant impact that a country's economy size and strength has on its population, understanding the phenomena that leads to the size dispersion becomes essential. One method used to understand interaction of variables is through mathematical modeling. Interestingly some models seem to fit for a variety of phenomena spanning diverse fields. In particular, Zipf's law, a power law, has become one of the most striking empirical facts in the social sciences and economics (Gabaix, 1999). Zipf's law has been found to be related to other known power laws that show up in as varied fields as the distribution of biological genera and species by Willis, size of cities by Auerbach, distribution of income by Pareto, and word usage frequencies by Zipf (Hill). Looking at the economic implications of previous research, the purpose of this project is to apply Zipf's law to the gross domestic product of nations in order to create a theoretical framework for the apparently vast differences between the few largest economies and the many smaller ones. From this the single research question used in this study follows: Can Zipf's law be used to model the rank and size of the gross domestic product of countries?

The remainder of this paper is organized as follows. Chapter II gives a background including definitions of key terms and the research plan. Chapter III shows the results of the research, particularly through the use of graphs demonstrating the data. Finally, Chapter IV gives a summary of the conclusions and suggestions for future research.

Background and Research Plan

Background

Zipf's Law. One of the most prominent regularities throughout the social science disciplines is that of Zipf's law for cities (Axtell and Florida, 2006). It says that the size distribution of cities within most countries seemingly fits a power law. Written mathematically, this means that "the probability that the size of a city is greater than some S is proportional to 1/S: $P(size>S)=\alpha/S^{\xi}$, with $\xi\approx1$" (Gabaix, 1999). If cities are ranked by size, creating a plot of the log of the size compared to the log of the rank, one would get a plot closely resembling a straight line with a slope of exactly -1 (Axtell and Florida, 2006). This law does not follow for some countries with unique social structures, such as China or the former USSR, but for other developed countries Zipf's law well approximates city size distribution (Marsili and Zhang, 1998). Although Zipf's law is most commonly used in reference to the rank and size of cities, its applications fall under a vast range of areas. In confirming Zipf's law's appearance within data, there are two necessary parts. First, evidence of the existence of a power law must be shown and second, the power law must have an exponent of -1 (Gabaix, 1999)

Zipf's law is so astounding because it seems collectively society organizes itself to follow this incredibly simple distribution law without the expressed desires of authorities (Marsili and Zhang, 1998).

Zipf's distribution essentially describes other phenomena including that of the distribution of firm sizes (Axtell and Florida, 2006).

In income distribution. Pareto in the late 19th century, described personal income as following a power law with an exponent of about 1.5, although after looking at several countries in 1922 Gini showed that income distributions can be estimated by power laws but with varying exponents (Okuyama, et. al., 1999). The Pareto distribution frequently referred to as the 80-20 law, suggests that approximately twenty percent of the population controls eighty percent of the wealth (Földvári, 2009). The Pareto distribution is often used to approximate Zipf's law (Meintanis, 2009).

Gibrat's law. Gibrat's law says that the mean and variance of the growth rate of an item are independent of its size (Hansberg, 2006). Gibrat's law has been studied in relation to areas such as financial returns, firms, and city sizes (González-Val & Sanso-Navarro, 2010).

Gibrat's law, or the law of proportionate effect as originally named by Robert Gibrat in 1931, is used to describe surprisingly non-random and extremely complex distributions (Eeckhout, 2004). Gabaix (1999) described an interaction between Zipf's law and Gibrat's law necessary in understanding distributions of gross domestic products used in this study.

Data Collection and Research Plan

Data Collection. For the purposes of this study, the data used will be gathered from the International Monetary Fund's World Economic Outlook Databases (World Economic and Financial Surveys, 2010). More specifically, the data used will be the gross domestic product of all 183 members of the International Monetary Fund from the years 2005 to 2008. The gross domestic product will be measured in terms of current prices in United States dollars.

This specific data was used for four main reasons. First of all, the International Monetary Fund had the most reliable and up to date data available during the time the research was completed. The database used is created biannually, beginning in January and June and appearing in publications in the following April and September.

Secondly, the data from the IMF from the years 2005 through 2008 was used for its completeness. The IMF provided data on all 183 countries counted within its membership for these years, with the exception of Turkmenistan in 2008 and estimates for a few countries. In the years preceding 2004, the gross domestic product numbers for multiple countries was incomplete and in the years after 2008 much more of the data were estimates done by the IMF.

Thirdly, the richness of the WEO database provided optimal data. The database provided not only the gross domestic products of the 183 countries but also provided further information about individual countries economies, specifically if the country had an emerging or advanced economy which proved beneficial in this study.

Finally, the data used was measured in terms of current prices and in United States dollars to provide a constant unit of measurement for the data. Using any variation would provide data that would not accurately provide growth rates and unison between countries necessary to properly analyze the data.

Research Plan. According to Gabaix (1999), to visualize Zipf's law, one orders the items being used, in this case countries, by rank (United States having the largest GDP has a rank of 1, Japan with the second largest GDP has a rank of 2, etc.). Next a graph is drawn with the log of the rank on one axis and the log of the GDP on the other. If Zipf's law follows, a straight line will appear. Furthermore the line will have a slope of -1.

In Gabaix's theoretical framework for explaining Zipf's law for cities, he uses a fixed number of items growing stochastically. Then assuming that for a particular range of sizes, the items follow Gibrat's law as defined above, he concludes that those particular items, in a steady state, will have a distribution that can be described by Zipf's law, including the power exponent of-1 (Gabaix, 1999).

With this in mind, the research plan was to follow a set of distinct steps for three different categories of data. First, the growth rates for each country were calculated individually. Secondly, the mean growth rate and standard deviation for the data were calculated in order to see if Gibrat's law holds for the data. Next the data were modeled by plotting the rank versus the GDP of the countries. This allowed for a visualization of the extreme gap and the obvious preponderance of data at the lower end of the scale. After this a second plot was created, plotting the log of the rank versus the log of each country's GDP in an attempt to visualize the linear behavior of the data consistent with a power law. If the data appeared linear, then the final step was to determine the power exponent by looking at the slope of the graph. If the slope is indeed -1, then Zipf's law is confirmed for the data.

After this procedure was done for all countries, the countries were split into two categories by their economies, advanced and emerging. The previous listed steps were repeated on each of these categories of data.

Results

All Economies

The International Monetary Fund provided information on a total of 183 countries. For the year 2008, information on Turkmenistan was not available. For each of the years, 2005 through 2008, the mean growth rate and the standard deviation from the mean are shown in the chart in Table 1.

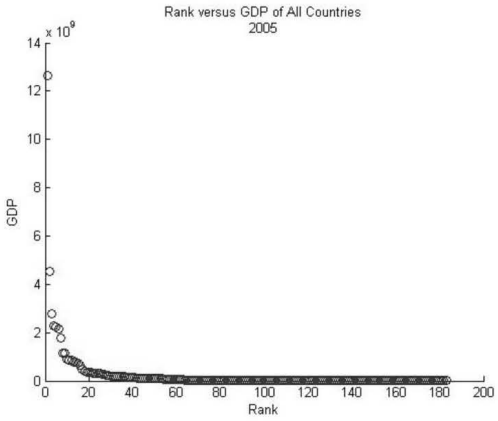
Table 1.

Years	Mean	Standard Deviation
2005 to 2006	0.140173176	0.098320774
2006 to 2007	0.183983363	0.092242753
2007 to 2008	0.176523999	0.090028962
2005 to 2008	0.605920054	0.362929085

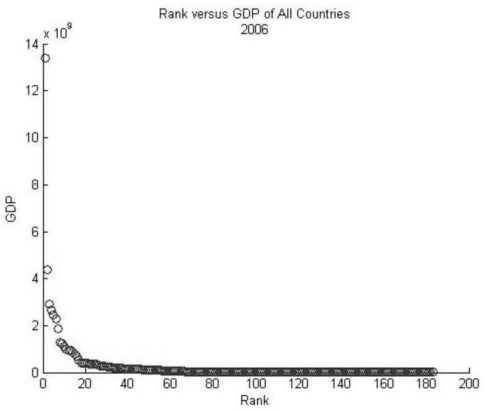
Next, the data from each of the four years was simply plotted on the graphs in Figure 1. Each of the GDPs from all 183 countries, with the exception of Turkmenistan in 2008, are plotted on the graphs. The clear separation among GDPs can be seen in these graphs. The number of small GDPs is obviously much more heavily weighted than the number of higher GDPs.

Figure 1.

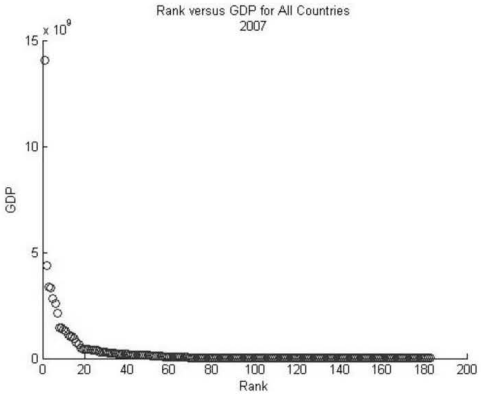
A.



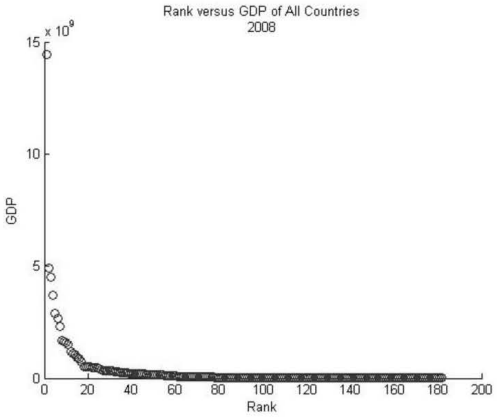
B.



C.



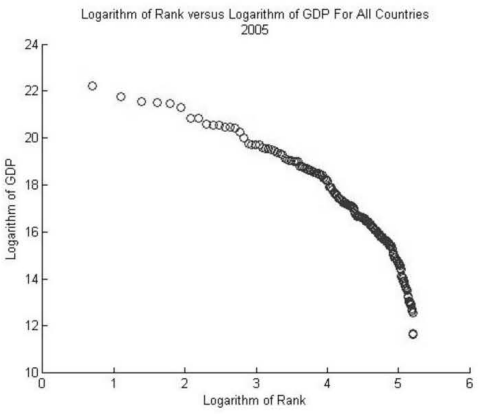
D.



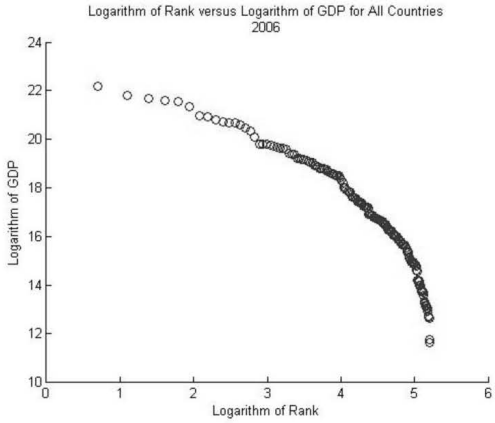
The next step in the research process was to graph the logs of rank versus the log of the GDPs for each of the 183 countries. The graphs created are shown in Figure 2. Clearly it can be seen immediately that these do not create a roughly linear graph, immediately exposing the absence of Zipf's law among all countries. Instead of a linear graph, the points create a distinctly concave down plot. Therefore the step of finding the slope and the power exponent is not necessary as neither Zipf's, nor the other power laws clearly apply.

Figure 2.

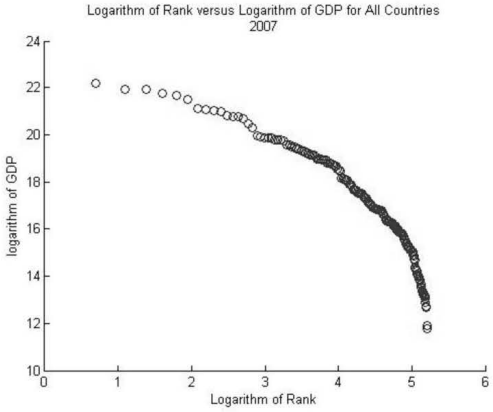
A.



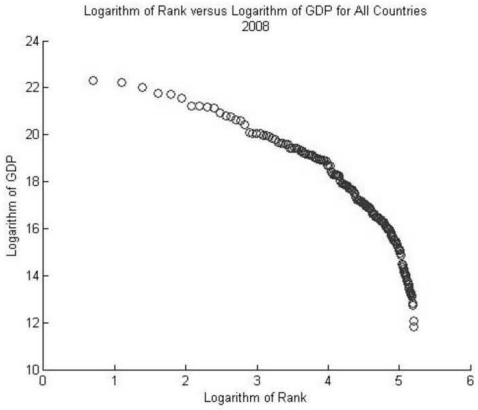
B.



C.



D.



Advanced Economies

According to the International Monetary Fund there are a total of thirty-three countries whose economies are seen as advanced. In continuing with the process used for all countries, the growth rates were calculated. Then, the average growth rate and the standard deviation from that growth rate were also calculated. The results from this calculation are shown in Table 2.

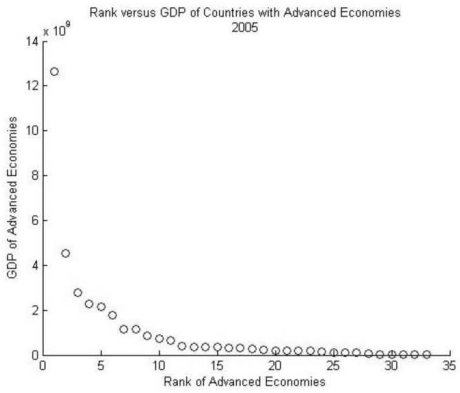
Table 2.

Years	Mean	Standard Deviation
2005 to 2006	0.045913326	0.07467498
2006 to 2007	0.15663896	0.062061399
2007 to 2008	0.090671167	0.088404904
2005 to 2008	0.360524724	0.192071891

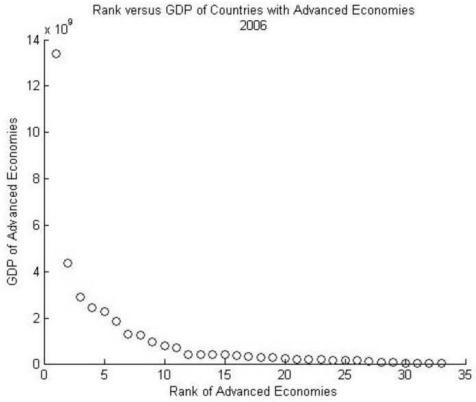
Next, as was previously done, the data from the thirty-three countries considered to have advanced economies by the International Monetary Fund were plotted and can be seen in Figure 3. Clearly among advanced countries as well, the graphs are much more heavily weighted in the tail end of the graph. With far fewer countries in this sample, each plotted point can be seen individually far more distinctly

Figure 3.

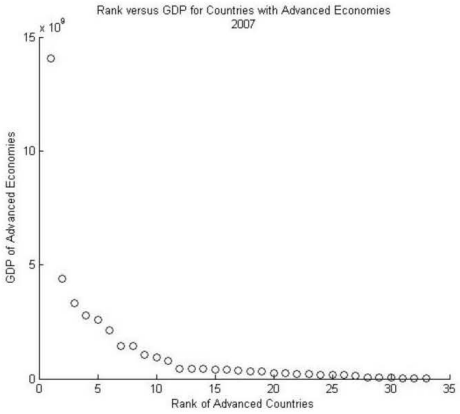
A.



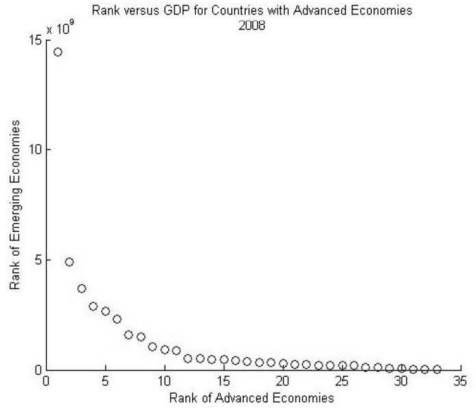
B.



C.



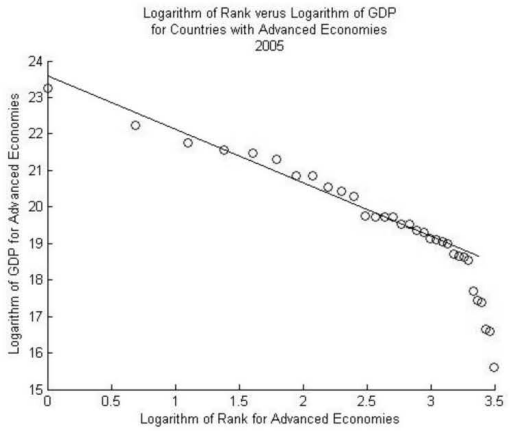
D.



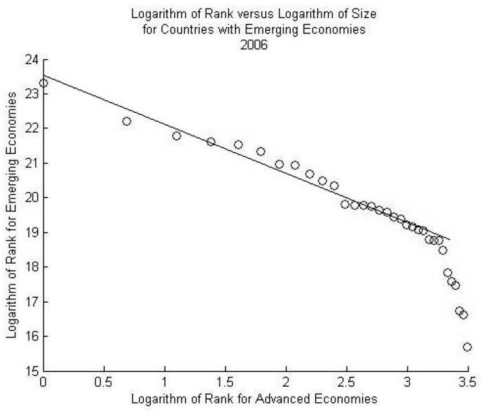
The third step in the process is to plot the logs of each of the ranks versus the logs of each of the GDPs for countries with advanced economies. The graphs are shown in Figure 4 with several fascinating things appearing. First of all, unlike the graphs of all the countries, the graphs of advanced economies do have a relatively linear distribution. On each of the graphs, the only exceptions to this linear design on the six lowest ranked countries. On each of the graphs, a line was placed approximating the linearity of the top ranked twenty-seven advanced countries.

Figure 4.

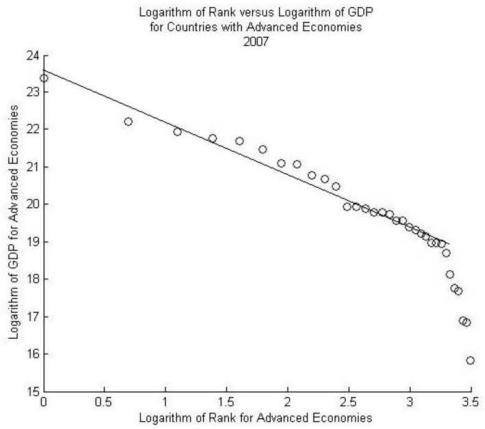
A.



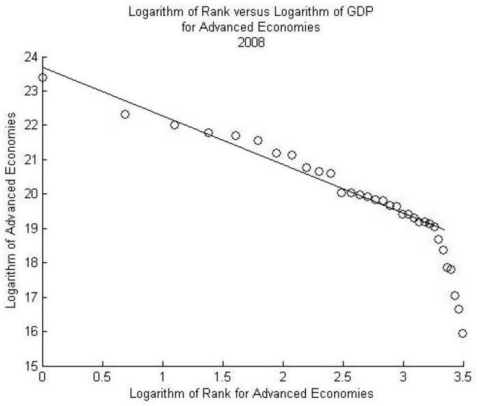
B.



C.



D.



With the linearity established for the logarithm plots of the advanced economies, the final step in the process is to find the slope of these lines and thus establishing the power exponent. For the year 2005, the exponent was determined to be -0.689737, for 2006 the exponent was determined to be -0.696829, for 2007, -0.714123, and for 2008, -0.72272.

Emerging Economies

Finally, after following the research steps for all countries and countries with advanced economies, the procedure was repeated a third time, using this time instead the remaining 150 countries with emerging economies as established by the International Monetary Fund. Once again, the growth rates were determined for each of the countries and then the mean and standard deviation of the growth rates over each year were determined. These results are displayed in Table 3.

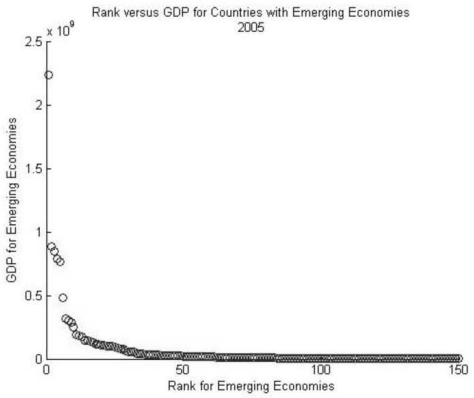
Table 3.

Years	Mean	Standard Deviation
2005 to 2006	0.15458278	0.10097654
2006 to 2007	0.18999925	0.096767624
2007 to 2008	0.19553834	0.113245655
2005 to 2008	0.66.0269356	0.368742368

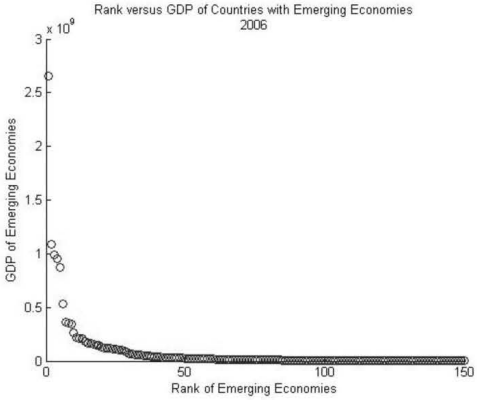
Next, as before the rank versus GDP for the 150 countries with emerging economies in order to display once again the much more heavily weighted lower end. Turkmenistan's economy fell into this section and so for the year 2008, only 149 countries' GDPs are plotted. These plots can be seen in Figure 5.

Figure 5.

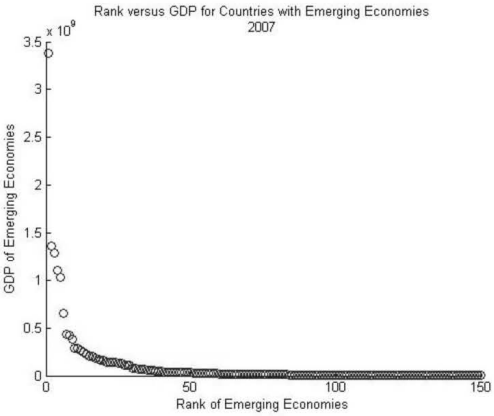
A.



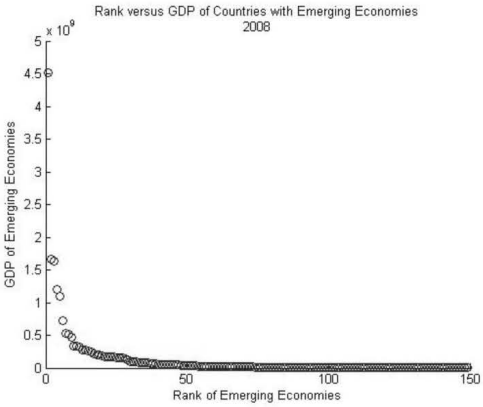
B.



C.



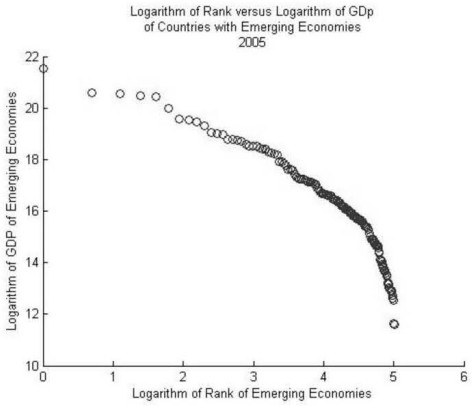
D.



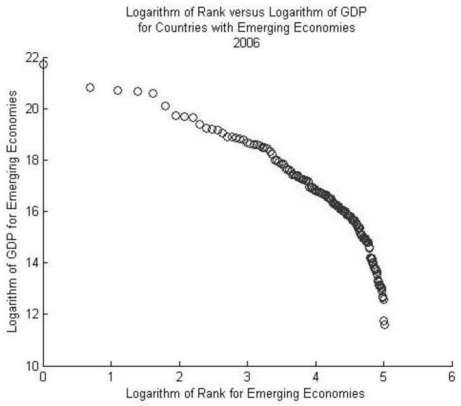
The next step, as was done for all countries and for countries with advanced economies, is to plot the logs of the ranks versus the logs of the GDPs of countries with emerging economies. These can be seen in Figure 6. As with the log plots of all countries, there is clearly once again not a linear distribution. Once again the graphs form a concave down distribution instead.

Figure 6.

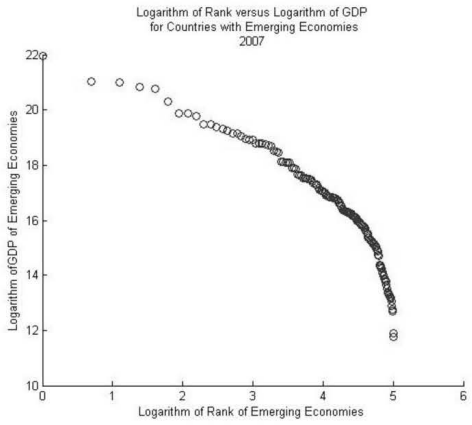
A.



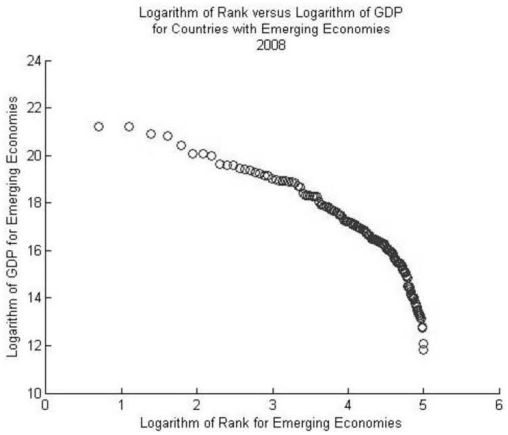
B.



C.



D.



Conclusions and Future Research

All Economies

In looking at the data shown in Figure 2, we can see that the log-log graphs for each of the four years for all economies all followed a similar pattern. In each of the four graphs, the data formed a roughly concave down pattern. As was mentioned previously in the results section, this means that all countries do not follow a power law and therefore the second step of determining the slope is not necessary to establish Zipf's law.

Although the explanation for this is not clear, a possible preliminary finding can be seen in Table 1. The standard deviation in growth rates among the 183 countries ranges between nine and ten percent between each pair of years from 2005 to 2008. Over the four year span the standard deviation was as high as 37 percent. Because of time constraints, it is difficult to determine whether this deviation is due to economy size. For this reason, it is quite clear that Gibrat's law also cannot be verified.

Advanced Economies

For advanced economies the results were somewhat different. In graphing the log of the rank versus the log of the GDP for each of the four years, a relatively linear pattern emerged. In each of the four graphs, all except approximately the lowest ranked six GDPs fell roughly on the same line. From this a power law was established. After the line estimating the data was added, the slopes for each of the four years were determined to be approximately between -0.70 and -0.72. Since the slope deviates significantly from -1.0, the conclusion based on this data is that the distribution of advanced economies does not follow Zipf's law but instead follows a Pareto distribution allowing for the variation in the power exponent. It is unclear why the lowest ranked economies seem to fall off the line.

A possible explanation for the Pareto distribution can be seen in Table 2. The standard deviation of growth rates between each pair of year ranges between approximately 7.5 and 9 percent while the standard deviation across all four years was approximately 19 percent, considerably lower than that of all economies. Time constraints on the project make it difficult to determine whether this deviation is independent of size making it unclear whether Gibrat's law was verified.

Emerging Economies

The log-log plots for GDPs from countries with emerging economies showed results very similar to those of all countries. Each of the four graphs showed a clearly concave down pattern and therefore does not represent a power law. Because of this, Zipf's law clearly does not apply to countries with emerging economies, making finding the slope irrelevant.

In looking at the growth rates as shown in Table 3, it can be seen that the standard deviation between each pair of years ranges roughly between 10 and 11.3 percent and nearly 37 percent across all four years. This is nearly double the standard deviation between growth rates of advanced economies suggesting a possible explanation for why neither Zipf's law, nor any other power law is demonstrated through the distribution of GDPs among emerging economies. Although the large deviation exists, because of time constraints it is difficult to determine whether this deviation is related to size. Although this deviation suggests this data does not follow Gibrat's law, it is not clear whether this is in fact the case.

Future Research

Several avenues exist to expand upon this research and upon the knowledge concerning power laws, and in particular Zipf's law. First, since the GDP of neither all economies nor emerging economies followed a power law, finding a model to describe these phenomena could prove valuable in understanding what causes this distribution. Secondly, since no clear cause for the power law distribution of advanced economies emerged, future research with fewer time constraints could be designed in order to determine two things, if the growth rates are independent of size as in Gibrat's law and as proposed by Gabaix (1999) and if not, what causes these economies to follow this pattern. A theoretical design could provide answers that this research was not able to provide. Finally, research can be done to continually expand upon the number of known distributions for which Zipf's law accurately models.

References

Axtell, R., Florida, R. (2006). Emergent cities: micro-foundations of Zipf's law.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *The American Economic Review*. 94 (5), 1429-1451.

Filmer, D., Pritchett, L. (1999). The impact of public spending on health: does money matter? *Social Science & Medicine*. 49, 1309-1323.

Földvári, P. (2009). Estimating income inequality from the tax data with a priori assumed income distributions in Hungary, 1928-41. *Historical Methods*. 42(3), 111-115.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*. 114(3), 739-767.

González-Val, R., Sanso-Navarro, M. (2010). Gibrat's law for countries. *Journal of Population Economics*, 23(4), 1371-1389. Doi:10.1007/s00148-009-0246-7

Haddad, L., Alderman, H., Appleton, S., Song, L., Yohannes, Y. (2003). Reducing child malnutrition: how far does income growth take us? *The World Bank Economic Review*, 17(1), 107-131. doi: 10.1093/wber/lhg012.

Hill, B.M. (n.d.). A theoretical derivation of the Zipf(Pareto) law. *University of Michigan*.

Marsili, M., Zhang, Y. (1998), Interacting individuals leading to Zipf's law. *Physical Review Letters*. 80(12), 2741-2744.

Meintanis, S. (2009), A unified approach of testing for discrete and continuous Pareto laws. *Statistics Papers Statisched Hefte*. 50(3), 569-580.

Okuyama, K., Takayasu, M., Takayasu, H., (1999). Zipf's law in income distribution of companies. *Physica A*.125-131.

Shen, C., Williamson, J.B. (2004). Accounting for cross-national differences in infant mortality decline (1965-1991) among less developed countries: effects of women's status, economic dependency, and state strength. *Social Indicators Research*. 53(3), 257-288. doi:10.1023/A:1007190612314

Rossi-Hansberg, E., Wright, M.L.J, (2007). Urban structure and growth. *Review of Economics Studies*. 74, 597-624.

World Economic and Financial Surveys (2010). *World Economic Outlook Database*.