Open Access

Corruption and Zipf's Law

Jang C. Jin[*,1], Bharat R. Hazari[2] and Thomas S. C. Lau[3]

[1]*Department of Decision Sciences and Managerial Economics, Chinese University of Hong Kong, Hong Kong*

[2]*Department of Economics and Finance, City University of Hong Kong, Hong Kong*

[3]*School of Accounting and Finance, Hong Kong Polytechnic University, Hong Kong*

Abstract: Zipf's law states that the population size of a city is inversely proportional to its population rank of the city. This paper examines the applicability of the Zipf's law to the world rank of corruption. The relationship between corruption and its rank is found to be approximately log-linear but less than perfect for Zipf's law. Due to a slight concavity of the relation, either a piecewise regression or a non-linear model provides an extremely convenient tool for predicting the degree of corruption across countries. Although limited number of observations, an alternative characterization of the corruption ranks appears to obey the Zipf's law more closely.

Keywords: Corruption, Zipf's law, log normal distribution.

1. INTRODUCTION

It is well established that many empirical distributions exhibit the properties of power laws. Zipf's [1] law is one of the most fascinating examples of such behavior in urban economics and geography. This law states that the population size of a city is inversely proportional to its population rank of the city. In contrast to the enormous literature in urban economics on Zipf's law in the form of rank-size regularity on city populations, there are few applications of this law to other areas of economics. Relationships similar to Zipf's law in economics have been originally documented by Pareto [2] and Gibrat [3] on distributions of income and firm sizes, respectively. Recent examples are Mantegna and Stanley [4] and Ulubasoglu and Hazari [5], among others. The former analyses the S&P index and the latter tourism.

In this paper we pose a question: Can Zipf's law be applied to corruption? Does corruption exhibit some sort of rank-size regularity? A positive answer to this question would allow us to predict the degree of corruption from a country's ranking. This would provide an excellent guide to estimating corruption without involving relative prices, income and poverty levels, forms of government, and so on, as explanatory variables. A more pressing question in this context is that if Zipf's law holds for corruption then what is its explanation? This is an open question for which many alternative explanations may be offered with corruption emerging from a random distribution that obeys some sort of power laws. Another motivation of this paper is to ascertain the use of corruption scores for the Zipf's law that relates ranks with frequencies. The frequency that the number of countries appears in each category of corruption ranks is further employed to check with the robustness of the results.

Our empirical results establish that Zipf's law strikes another area of economics--corruption. The regression results show that a linear fit on the data for 163 countries explains 87% of the variations in corruption although regression coefficients are less than perfect for the Zipf's law. A closer examination of scatter diagrams exhibits a concave rank-size plot; once we take into account of this concavity, the regression results explain nearly 99% of the variations. Corruption is thus generated from a log normal distribution that gives rise to a slight concavity. The concavity, however, mitigates substantially when a frequency (the number of countries) that appears in each category of corruption ranks is used. Note, however, that we have not followed the practice of truncating the sample to generate a better fit.

Before presenting our results we would like to make some comments on corruption. The Merriam-Webster dictionary defines corruption in the following manner:

> *Corruption: 1a: impairment of integrity, virtue, or moral principle: DEPRAVITY b: DECAY. DECOMPOSITION c: inducement to wrong by improper or unlawful means (as bribery) d: a departure from the original or from what is pure or correct.*

Taken from Merriam-Webster online

This is an extraordinarily wide ranging definition. It encompasses many areas of human life and does not specifically pass corruption as belonging to the domain of economics. In a very broad sense, corruption also begins within a family. The denial of human rights is an example of corruption, so sending children into workforce is a corrupt behavior. Similarly feticide is also a corrupt behavior. Such behavior needs not be necessarily related to economic

[*]Address correspondence to this author at the Department of Decision Sciences and Managerial Economics, Chinese University of Hong Kong, Hong Kong; Tel: (852) 2609-7902; Fax: (852) 2603-5104;
E-mail: jcjin@cuhk.edu.hk

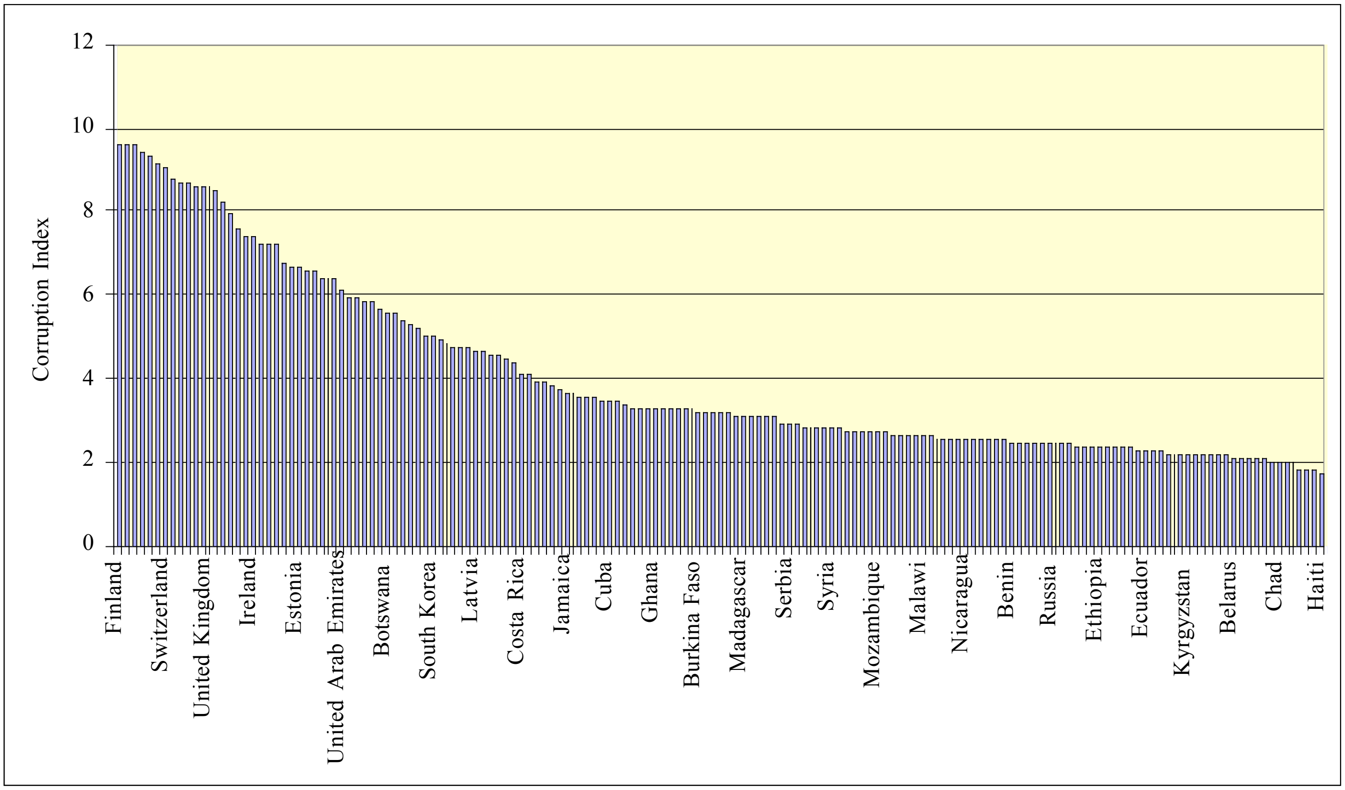

Fig. (1). World Corruption Index, 2006. Source: Transparency International (2006). Total 163 countries were included in a graph but country names in the horizontal axis appear only in every 6th ranking.

phenomenon which would identify corruption with functioning of markets, relative prices, income levels, and so on. We believe that it is the randomness in the occurrence of corruption and country specific forces (for example, preference for boys over girls, sale of children to prostitution) that give rise to Zipf's law in corruption.

2. BASIC RESULTS

Fig. (**1**) shows the world corruption index for the year 2006. The corruption index ranks 163 countries in terms of the degree to which corruption is 'perceived' to exist among public officials and politicians [6]. The perceived corruption scores are the weighted averages of several corruption data independently surveyed by varied research institutions all over the world, and the country with the highest score is the one perceived to be least corrupt. For example, Finland, Iceland, and New Zealand rank number one with a score of 9.6 out of 10, which means that the three countries are perceived to be most transparent in public services in the world. In general, many developed countries appear in this top tier. Relatively small but rich countries like Switzerland also enter the higher ranks. After that, the corruption scores gradually fall. For low income countries, the high degree of corruption (i.e., low corruption indexes in our graph) appears to be similar to each other. The most corrupted countries in the sample are Iraq, Guinea, Myanmar, and Haiti with all less than 2.0 in corruption scores.

Fig. (**2**) shows the line fit of a reciprocal model that explains a salient feature of the Zipf's law. For example, the corruption index in the vertical axis may occur as an inverse of the rank measured in the horizontal axis. That is, f = φ (1/r), where f = frequency occurred, and r = ranks. Suppose that φ = 10 initially as a maximum score of perceived corruption. Then, f = 10 for the rank number one, f = 5 for the rank number two, f = 3.33 for the third rank, and so forth. An attempt to fit the reciprocal model using the corruption data gives the following results.

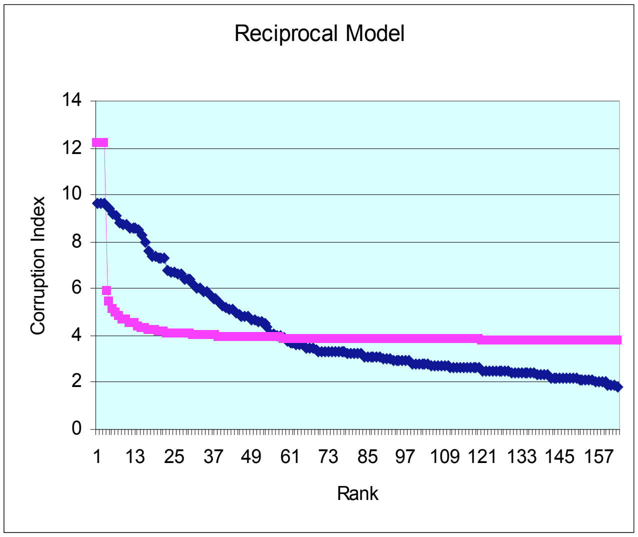

Fig. (2). Line fit of reciprocal model. Note: A dark blue line represents actual data, while a light pink color represents a predicted line.

$$CI_i = 3.73 + 8.50\ (1/\text{Rank}_i) \qquad (1)$$

$$(1.05)^*$$

$$R^2 = 0.288 \qquad \sigma = 1.822 \qquad n = 163$$

where CI_i is the corruption index (10 for least corrupt, and 0 for most corrupt), where i = 1, 2, …, 163 countries, and Rank_i represents the corruption rank for country i. The number in parenthesis represents standard error of the parameter estimate. The slope coefficient is statistically significant at the conventional significance levels, yet the estimated R^2 is rather low. The estimated regression line (light pink color) is depicted together with actual data (dark blue color) in Fig. (**2**). Large deviations are observed between actual and predicted values although the parameter estimate appears to be significantly different from zero at the 5% significance level. The standard error estimate of the model (σ) appears to be relatively large because the raw data were used in levels for estimation of the model (1).

Fig. (**3**) then plots the corruption index against the corruption rank in logarithms. The corruption rank is assigned, based on corruption scores, from the least corrupt to the most corrupt countries. In general, rich countries are found in higher ranks (least corrupt), while poor countries appear in lower ranks (more corrupt). More specifically, the corruption ranks are highly correlated with per capita income (approximately r = -0.8). Top-20 least corrupt countries are the ones with per capita GDP $20,000-45,000; next twenty least-corrupt countries are within the income range between $10,000 and $20,000; and so forth. At last, thirty most-corrupt countries in the bottom ranks have per capita income less than $1,000.

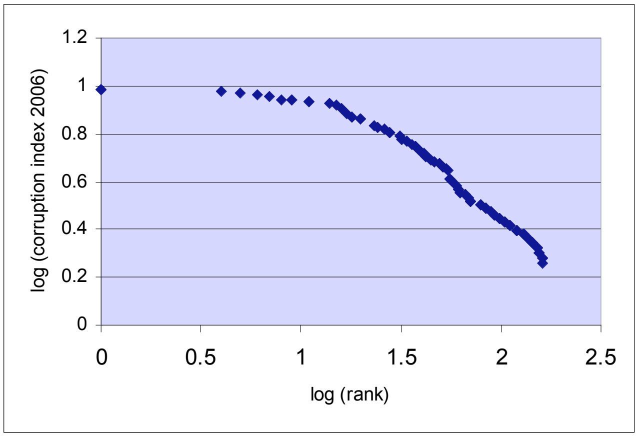

Fig. (3). Log corruption *vs* log rank. Source: Transparency International (2006).

An interesting question that arises is whether these graphs obey Zipf's Law. In other words, is there a linear relationship between the log corruption and the log rank? If the relationship is exactly linear, it will follow the empirical law found in Zipf [1]. Thus, a perfect case of the Zipf's law would be the one that the regression coefficient of a log-linear model is a minus one, since the corruption index may follow φ (1/rank), where the maximum corruption index φ = 10. In other words, log (corruption) = 1 – log (rank). However, as in Fig. (**3**), the relation between the two variables may not be linear because a diminishing speed of the corruption index in higher ranks particularly from the 1st to the 16th is slower than the one expected in the Zipf's law. For example, for the three rank-number-one countries (i.e., log 1 = 0 on the horizontal axis), the vertical intercept in Fig. (**3**) is close to 1 (more precisely, log 9.6 = 0.98); after that, the log corruption falls slowly because the corruption scores are similar to each other in higher ranks. In contrast, for the corruption ranks lower than the 16th, i.e., log (rank) > 1.2, the negative association between the two variables seems to be approximately linear.

The linear relationship is further investigated using a simple linear regression model:

$$\log (CI)_i = \beta_0 + \beta_1 \log (\text{Rank})_i + \varepsilon_{i.} \qquad (2)$$

where ε_i is white noise residuals. The dependent and independent variables are taken as logarithms and thus a heteroscedasticity problem — residuals are relatively large in higher-rank countries and smaller in lower ranks — may not be serious in this case because the log-linear model normally mitigates the measurement scale of the raw data. The first equation in Table **1** uses the corruption index of the year 2006 as a dependent variable and provides evidence that the linear relationship is explained approximately 87% although the regression coefficient appears to be less than perfect for the Zipf's law. Standard error estimates (σ) are relatively small in this log-linear model. The results are, in general, consistent with the Zipf's law in which the log of corruption in the vertical axis is approximately linearly related with the log of ranks measured in the horizontal axis.

The second model in Table **1** uses a four-year average of the corruption indexes over the period 2003-2006. For the robustness of the results, the third model employs another four-year average over earlier years 2000-2003. The last one uses a seven-year average over the entire sample periods 2000-2006. For all different measurements of the corruption index, little variation is found in parameter estimates, as well as in R^2. The results are generally robust across different measurements of the corruption index. This also indicates that people's perception on a country's degree of corruption changes little over time.

Fig. (**4**) plots the line fits that correspond to the four regression results discussed in Table **1**. As noted earlier, the predicted regression lines do not perfectly fit the actual data. Large error estimates are observed in higher ranks, perhaps due to small differences in least-corrupt countries. A piecewise linear regression model will fit the data even better if the regression lines are divided into two different parts based upon the degree of corruption.

3. ALTERNATIVE REGRESSION MODELS

3.1. Piecewise Linear Regression

A careful examination of Fig. (**4**) shows that there is one linear relationship up to a certain rank and another one after that rank. More specifically, the corruption index falls approximately linearly until the threshold rank 16, after which it also decreases linearly but at a much steeper rate. Thus, we construct and run a piecewise linear regression that consists of two linear segments. A threshold rank is assumed

Table 1. Regression Results

Model: log (corruption index) = $\beta_0 + \beta_1$ log (corruption rank) + ε_i

Corruption Index	β_0	β_1	obs	R^2	σ
Corruption index year 2006	1.342	-.441(.014)	163	.868	.073
Average index 2003-2006	1.400	-.485(.016)	132	.874	.075
Average index 2000-2003	1.374	-.491(.028)	82	.789	.101
Average index 2000-2006	1.368	-.484(.027)	82	.798	.097

Note: Standard errors are in parentheses.

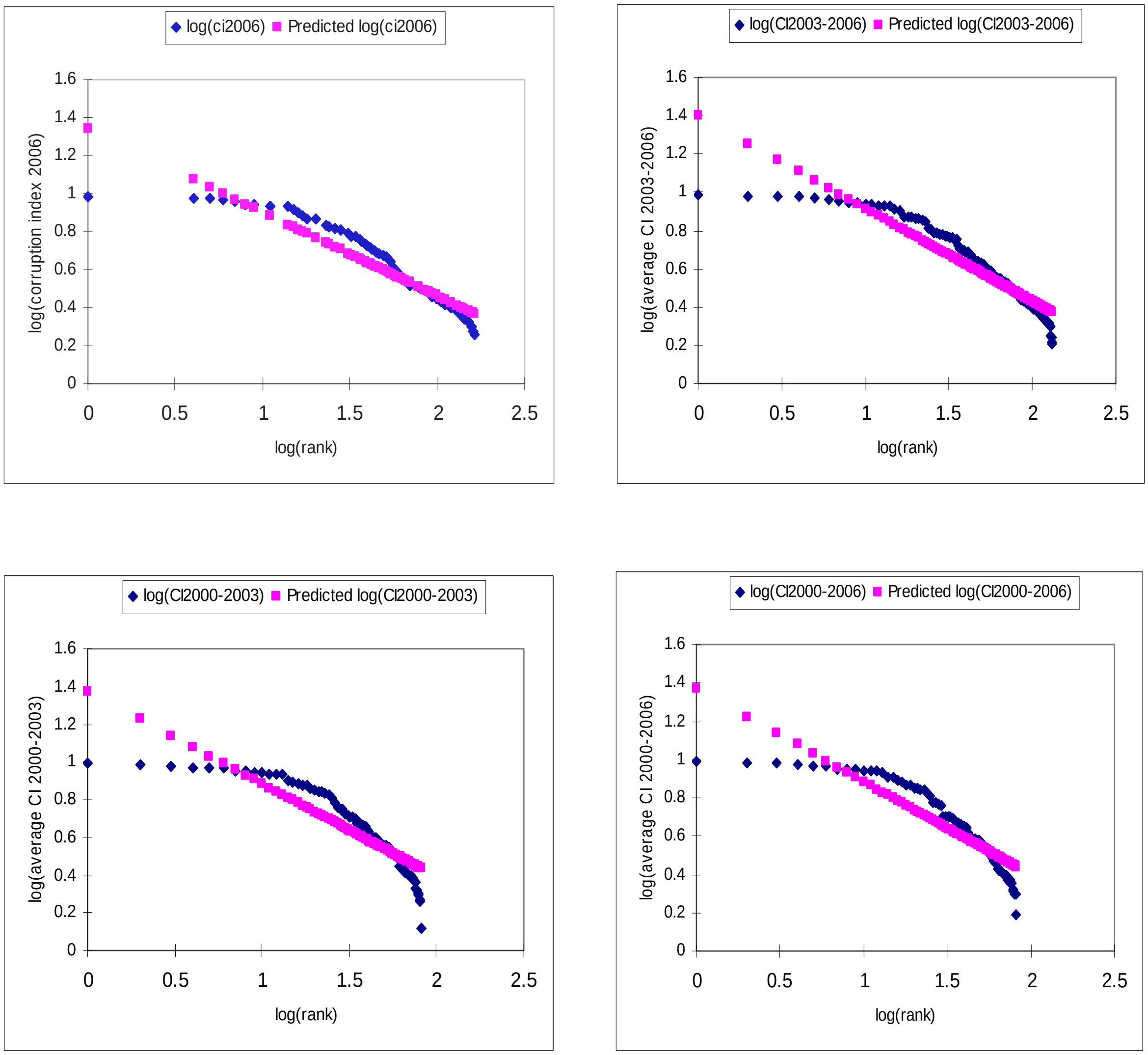

Fig. (4). Line fit plots.

to be the 16th, and the technique of dummy variables is used to estimate the two different slopes of the two segments within a model. We obtain the following results.

$$\log CI_i = .981 - \underset{(.009)^*}{.029} \log Rank_i - \underset{(.012)^*}{.605}\, D_i (\log Rank_i - 1.2041) \quad (3)$$

$$R^2 = 0.992 \qquad \sigma = 0.0178 \qquad n = 163$$

where a dummy variable $D_i = 1$ if log $Rank_i > 1.2041$ (i.e., log 16 = 1.2041) and 0 otherwise. The values in parentheses are the estimated standard errors. This piecewise or spline regression improves the fit of the model to 99%, and standard error estimates are significantly reduced to 0.0178. Parameter estimates show correct signs and they are all statistically significant at the conventional significance

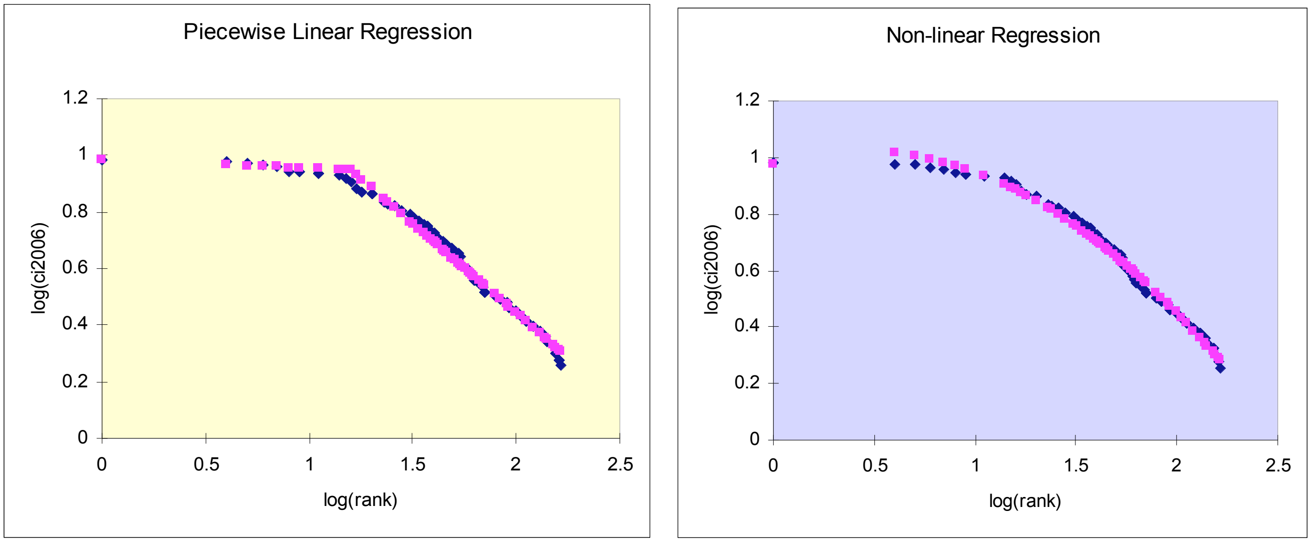

Fig. (5). Alternative regression models.

levels. The slope of the first segment is -0.029, while the slope of the second segment is -0.634. The difference between the two slopes (-0.605) appears to be statistically significant at the 5% significance level. The two different slopes are portrayed in the first panel of Fig. (**5**).

3.2. Non-Linear Regression

To capture the non-linearity displayed in Fig. (**4**), we further estimate a non-linear regression model that includes the square of the log rank; that is, instead of using a linear form, the relation between the log corruption and the log rank is now quadratic.

$$\log CI_i = .974 + .209 \log Rank_i - .236 (\log Rank_i)^2 \qquad (4)$$

$$(.014)^* \qquad (.005)^*$$

$$R^2 = 0.992 \qquad \sigma = 0.0186 \qquad n = 163$$

The estimated coefficient of the linear term appears to be significantly positive, while the estimated square term is negative and significant. The negative coefficient on the square term indicates that the force toward the positive relation between the log corruption and the log rank mitigates as the rank falls. In other words, the relations are initially positive for higher ranks but become negative for lower ranks. The positive coefficient of the linear term found in (4) is due to a nearly flat and small variation in higher ranks. Once we dropped the first 16 countries in higher ranks from the sample, all coefficients were found to be significantly negative for both linear and non-linear terms (-0.24 and -0.11, respectively).[1] Therefore, if we move from least-corrupt to more-corrupt countries, the log corruption index decreases slowly in higher ranks but it decreases at an increasing rate in more corrupt countries. This implies a slight concavity of the non-linear relation, as shown in the second panel of Fig. (**5**).

4. FURTHER CHARACTERIZATION OF CORRUPTION RANKS

The findings of concave non-linear relationships between corruption and its rank have used the original corruption index that assigned 0s to most corrupt countries and 10s to least corrupt countries. Alternatively, corruption ranks are re-characterized as number one (most corrupt countries) if the corruption scores are less than or equal to 2 points. The corruption rank goes down further to number two, number three, and so on if less corrupted. The bottom rank is 8th place with the corruption scores that are between 9 to 10 points (least corrupt countries). For each category of corruption ranks, the number of countries included is counted as a frequency. This would be another way of ranking corruption.[2] In this way, we find that there are a large number of high-corrupt countries and very few low-corrupt ones, a distribution similar to Zipf's law.

To examine the Zipf-like version, we further estimate the log-linear model that uses the new ranking of corruption. The result is obtained as follows.

$$\log Freq_i = 1.811 - 1.15 \log Rank_i \qquad (5)$$

$$(0.12)^*$$

$$R^2 = 0.93 \qquad \sigma = 0.10$$

where $Freq_i$ is the number of countries in each category of corruption ranks, and $Rank_i$ represents the new ranking of corruption (i.e., 1 for most corrupted countries and 8 for least corrupted ones). The number in parenthesis represents standard error of the parameter estimate. We find a nearly perfect fit, as is the case of Zipf's. The slope coefficient appears to be close to minus one and statistically significant at the conventional significance levels. Standard error estimates are also very small. The results are in general

[1] The results are available upon request.

[2] This method was suggested by an anonymous referee of this *journal*.

consistent with the Zipf's law in which the log of a frequency in the vertical axis is approximately linearly related to the log of corruption ranks measured in the horizontal axis. Fig. (**6**) plots the line fit that corresponds to the regression result in equation (5). The predicted regression line nearly perfectly fits the actual data, except for a few observations in lower ranks. Other than that, the association appears to be approximately linear. Therefore, the frequency-based corruption ranks are broadly consistent with Zipf's law.

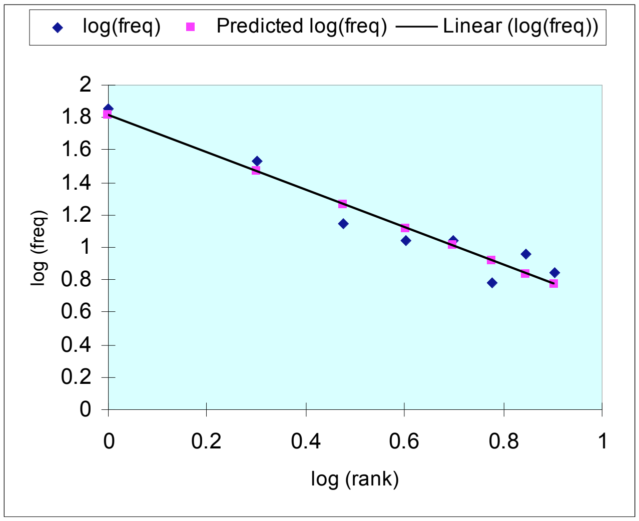

Fig. (6). Further characterization of corruption ranks.

5. CONCLUSION

A key finding from the regression results is that Zipf's law holds approximately for the world rank of corruption. Estimation of log-linear models provides evidence that a negative relationship between the log of corruption and the log of ranks is approximately linear although the regression coefficient appears less than perfect for the Zipf's law. The model fit is improved to 99% with the use of a piecewise regression model, and a non-linear regression model that includes a quadratic term also shows a slight concavity of the relation between log corruption and log ranks. The log normal distributions are mainly due to a diminishing speed of the corruption index that differs across countries. The concavity of the relation, however, mitigates significantly when a frequency (the number of countries) that appears in each category of corruption ranks is used.

While there may well be other interpretations of this relationship between corruption and its rank, one explanation for this finding would be the Zipf's law that strikes another area of economics -- corruption. The degree of corruption is highly correlated with per capita income. Most-corrupt countries in bottom ranks have per capita income less than $1,000, whereas least-corrupt countries are in general rich and developed economies. Developed countries also have a much higher level of education which in principle may reflect negative attitudes against corruption. However, there are very few low-corrupt countries in the world; most developing countries and many underdeveloped countries are highly corrupted. The shape of the distribution is a hyperbola type, which lends support to the Zipf's law regarding the rank-size regularity on corruption.

REFERENCES

[1] Zipf G. Human behavior and the principle of least effort, Cambridge, MA: Addison-Wesley 1949.

[2] Pareto V. Cours d'economie politique. Geneva, Switzerland 1896.

[3] Gibrat R. Les inegalites economiques. Paris, France: Libraire du Recveil Sirey 1931.

[4] Mantegna RN, Stanley HE. The scaling behavior of an economic index. Nature 1995; 376(6535): 46-9.

[5] Ulubasoglu MA, Hazari BR. Zipf's law strikes again: the case of tourism. J Econ Geogr 2004; 4(4): 459-72.

[6] Transparency International. Corruption Perception Index 2006. Available at: http://www.transparency.org

Received: October 23, 2009 Revised: March 6, 2010 Accepted: May 14, 2010