



Testing for Zipf's Law: A Common Pitfall

Carlos M. Urzúa*

Documento de Trabajo
Working Paper

EGAP-2010-04

Tecnológico de Monterrey, Campus Ciudad de México

*EGAP, Calle del Puente 222, Col. Ejidos de Huipulco, 14380 Tlalpan, México, DF, MÉXICO
E-mail: curzua@itesm.mx

Testing for Zipf's Law: A Common Pitfall

Carlos M. Urzúa*
December 2010

Abstract

It is noted that the regression procedure commonly used when testing for Zipf's law is erroneous.

JEL Classification: C12, R11, R12

Keywords: Zipf's law, rank-size.

Choose n cities within a country and rank them by size to get the ordered sequence $x_{(1)} \geq \dots \geq x_{(r)} \geq \dots \geq x_{(n)}$. Zipf's law, also known as the rank-size relation, asserts that a graph of the rank against the size would then render a rectangular hyperbola. That is,

$$rx_{(r)} = c . \quad (1)$$

for a constant c and all r . Zipf (1949) claimed that this relation holds for many different sorts of objects, ranging from the frequencies of natural language utterances to the number of species in biological genera (see Li, 2002, for a review). Although that asseveration has been disproved in most instances since then, it continues to be invoked in some cases, such as the size of cities.

Unfortunately, most of the authors that defend Zipf's claim use a testing procedure that is erroneous: They estimate, through ordinary least squares (OLS), the regression

$$\ln(r) = \beta_1 + \beta_2 \ln(x_{(r)}) + \varepsilon_r \quad (2)$$

for $r = 1, \dots, n$, and then test the null hypothesis that $\beta_2 = -1$. Aside from the fact that the OLS estimators are not efficient in that case, given that r is an integer, what makes (2) not only dubious but plainly wrong is the fact that the intercept is not a nuisance parameter in the

*Tecnológico de Monterrey, Campus Ciudad de México, Calle del Puente 222, Colonia Ejidos de Huipulco, 14380 Tlalpan, México, DF, México (e-mail: curzua@itesm.mx).

regression. This is so because in terms of, say, the largest or the smallest objects, equation (1) implies that $c = x_{(1)}$ or $c = nx_{(n)}$. Alperovich (1984) uses the first expression, but, as seen below, the second one has to be preferred to the other $n-1$ alternatives. Thus, if one is inclined to use a regression model to test for Zipf's law (something that we do not recommend), the correct procedure would be to estimate

$$\ln(r) = \theta \ln(x_{(r)} / nx_{(n)}) + \varepsilon_r, \quad (3)$$

for $r = 1, \dots, n-1$, and then test the null hypothesis that $\theta = -1$.

Such an ad-hoc test can be replaced with more formal ones after uncovering the probability law that is behind the rank-size relation (wrongly called the rank-size distribution!). As shown by, e.g., Rapoport (1978) or Urzúa (2000), the probability law behind (1) corresponds to the Paretian density function $f(x) \propto (x/\mu)^{-2}$, where $x \geq \mu > 0$ (the first of these last inequalities is the reason for having chosen $c = nx_{(n)}$ above). Among a number of parametric and nonparametric tests for that distribution, Urzúa (2000) proposes in particular the following simple test statistic:

$$LMZ = 4n[z_1^2 + 6z_1z_2 + 12z_2^2], \text{ where } z_1 = 1 - \frac{1}{n} \sum_{i=1}^n \ln \frac{x_{(i)}}{x_{(n)}}, \quad z_2 = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n \frac{x_{(n)}}{x_{(i)}}, \quad (4)$$

which is asymptotically distributed under the null as a chi-square with 2 degrees of freedom. An appealing feature of the test is that it is locally optimal if the alternatives are other power laws.

In order to illustrate the use of (3) and (4), consider the first results reported by Rose (2006, Table 1) in his interesting paper on the size distributions of, both, cities and countries across the world. Using regressions like (2), and choosing 5% as the level of significance, he cannot reject Zipf's law in particular for the case of the 50 largest US combined statistical areas in 1990 and 2000 (CSA19 and CSA20), as well as in the case of the 200 largest US metropolitan

and micropolitan statistical areas, also in 1990 and 2000 (MSA19 and MSA20). Nevertheless, if one uses (3), the following slope estimates (and robust standard errors) are found for CSA19, CSA20, MSA19 and MSA20: -1.046 (.008), -1.040 (.009), -1.035 (.004) and -1.032 (.003). Thus, Zipf's law is handily rejected in the four cases with that level of significance. Alternatively, after computing (4) the following LMZ values are obtained: 2.881, 2.097, 10.732, and 8.728, respectively. Thus, the null hypothesis is rejected in the case of the MSAs for both years.

As a final comment we note that in other sciences, such as Physics, Zipf's law is meant to encompass the more general rank-size relation $rx_{(r)}^\alpha = c$, with $\alpha > 0$. In that case the points made here can be readily generalized. In particular, the probability law behind that general relation can be shown to be Pareto's law, which has a density function of the form $f(x) \propto (x/\mu)^{-(\alpha+1)}$, $x \geq \mu > 0$. Needless to add, several parametric and non-parametric tests are already available in the literature for that distribution.

REFERENCES

- Alperovich, Gerson (1984). "The Size Distribution of Cities: On the Empirical Validity of the Rank-Size Rule." *Journal of Urban Economics* 16, 232-239.
- Li, Wentian (2002). "Zipf's Law Everywhere." *Glottometrics* 5, 14-21.
- Rapoport, Anatol (1978). "Rank-Size Relations." In *International Encyclopedia of Statistics*, vol. 2, edited by William H. Kruskal and Judith M. Tanur, pp. 847-854. New York: Free Press.
- Rose, Andrew K. (2006). "Cities and Countries." *Journal of Money, Credit and Banking* 38, 2225-2245.
- Urzúa, Carlos M. (2000). "A Simple and Efficient Test for Zipf's Law." *Economics Letters* 66, 257-260.
- Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA.: Addison-Wesley.