SOFTWARE PIRACY AND HOW IT IS INFLUENCED BY THE CORRUPTION LEVEL FOR ANY GIVEN COUNTRY. OPEN SOURCE AND FREE SOFTWARE AS SOLUTIONS TO THE PROBLEM.

Dragos-Paul Pop[1]

Abstract

Today's IT world is slowly driving towards open source and open software trends. Even Microsoft is taking such approach with some of its software products (the MSDNAA program is the best example). Although everyone is happy that software is becoming cheaper or even open source, we must ask ourselves what led to this trend. Why are software companies giving out software products for free when just a few years ago they were charging us big money for it? The answer is, of course, marketing issues. But another big factor is the piracy factor.

Keywords: software piracy, corruption, open source, free software, statistic, test

Introduction

What is software piracy or digital piracy? Piracy is the act of distributing something without the given consent of the author of that product. In other words, it is illegal to share your copy of the software product with anyone else, because everybody should buy their own copy. Buying software insures that the people that are building it are getting paid and they can continue to develop. It seems only fair. But software is very easy to duplicate and distribute. Of course, there are a lot anti-piracy methods like serial numbers and internet activation, but because software is, in essence, just a big collection of algorithms and instructions, it can very easily be decoded and those anti-piracy methods removed or counteracted. The question is why would people want to obtain their software from illegal sources? The answer is simple: because people tend to choose what's cheaper and don't really care about anyone else.

This paper tries to prove that software piracy is closely related to a big society issue: corruption.

The simple regression model can be used to see if the Corruption Perceptions Index of a given country influences the rate of software piracy in that country. We will try to see that if a country is seen as corrupt that will lead to the rise in software piracy, because it's citizens will obey the digital copyright law less than the citizens of a less corrupt country.

Data sources

This paper gathered it's data from two international sources: Transparency International and Business Software Alliance:

[1] Assistant teacher at the Romanian-American University in Bucharest. Ph.d. candidate at the Academy of Economic Studies in Bucharest. E-mai: dragos_paul_pop@yahoo.com

- Corruption Perceptions Index: http://www.transparency.org/policy_research/surveys_indices/cpi/2009/cpi_2009_table
- Software Piracy Rate : http://portal.bsa.org/globalpiracy2009/index.html http://portal.bsa.org/globalpiracy2009/studies/09_Piracy_Study_Report_A4_final_111010.pdf

The simple regression model defined and used:

To define the model, we will use the following annotations:

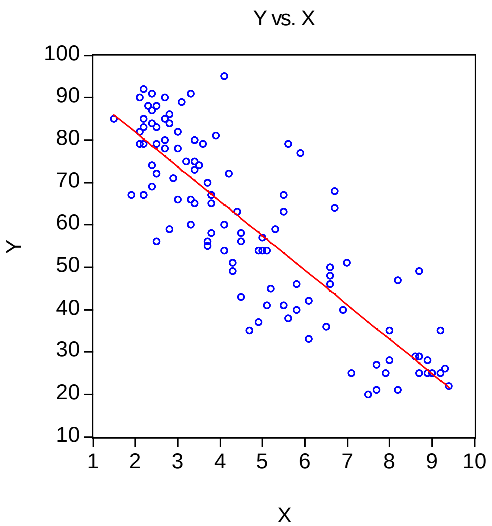
- x = corruption perceptions index
- y = software piracy level

The model becomes:

y = f(x)+e

Because the empiric points graph shows that the distribution can be approximated using a straight line, the model becomes:

$y_t = a + bx_t + \varepsilon_t$; t = 1..110



With the significance of the two variables in mind, we can make the following statements:

- parameter ***a*** represents an autonomous part of the software piracy percentage because for x = 0 we have y = a;
- parameter ***b*** represents the slope of the line or the regression coefficient for the software piracy percentage

Identifying the data series needed to estimate the parameters of the regression model

Country	y	x			
United States	20	7.5	Croatia	54	4.1
Japan	21	7.7	Lithuania	54	4.9
Luxembourg	21	8.2	Poland	54	5.0
New Zealand	22	9.4	Colombia	55	3.7
Australia	25	8.7	Brazil	56	3.7
Austria	25	7.9	Latvia	56	4.5
Belgium	25	7.1	Mauritius	56	2.5
Finland	25	8.9	Jordan	57	5.0
Sweden	25	9.2	Greece	58	3.8
Switzerland	25	9.0	Malaysia	58	4.5
Denmark	26	9.3	Costa Rica	59	5.3
United Kingdom	27	7.7	Egypt	59	2.8
Germany	28	8.0	Kuwait	60	4.1
Netherlands	28	8.9	Mexico	60	3.3
Canada	29	8.7	Oman	63	5.5
Norway	29	8.6	Turkey	63	4.4
Israel	33	6.1	Chile	64	6.7
Ireland	35	8.0	India	65	3.4
Singapore	35	9.2	Romania	65	3.8
South Africa	35	4.7	Bosnia and Herzegovina	66	3.0
United Arab Emirates	36	6.5	Morocco	66	3.3
Czech Republic	37	4.9	Brunei	67	5.5
Taiwan	38	5.6	Bulgaria	67	3.8
France	40	6.9	Ecuador	67	2.2
Portugal	40	5.8	FYROM (Republic of Macedonia)	67	3.8
Hungary	41	5.1	Russia	67	2.2
South Korea	41	5.5	Venezuela	67	1.9
Spain	42	6.1	Uruguay	68	6.7
Slovakia	43	4.5	Philippines	69	2.4
Malta	45	5.2	Peru	70	3.7
Puerto Rico	46	5.8	Argentina	71	2.9
Slovenia	46	6.6	Lebanon	72	2.5
Hong Kong	47	8.2	Tunisia	72	4.2
Cyprus	48	6.6	Panama	73	3.4
Iceland	49	8.7	Honduras	74	2.4
Italy	49	4.3	Serbia	74	3.5
Estonia	50	6.6	Albania	75	3.2
Qatar	51	7.0	Thailand	75	3.4
Saudi Arabia	51	4.3	Dominican Republic	77	5.9
Bahrain	54	5.1	Kazakhstan	78	2.7

Senegal	78	3.0	**Iraq**	85	1.5
Botswana	79	5.6	**Ukraine**	85	2.2
China	79	3.6	**Vietnam**	85	2.7
Ivory Coast	79	2.1	**Indonesia**	86	2.8
Kenya	79	2.2	**Belarus**	87	2.4
Nicaragua	79	2.5	**Azerbaijan**	88	2.3
Bolivia	80	2.7	**Libya**	88	2.5
El Salvador	80	3.4	**Sri Lanka**	89	3.1
Guatemala	80	3.4	**Armenia**	90	2.7
Montenegro	81	3.9	**Yemen**	90	2.1
Paraguay	82	2.1	**Bangladesh**	91	2.4
Zambia	82	3.0	**Moldova**	91	3.3
Cameroon	83	2.2	**Zimbabwe**	92	2.2
Nigeria	83	2.5	**Georgia**	95	4.1
Algeria	84	2.8			
Pakistan	84	2.4			

Descriptive analysis of the data series

This analysis is done using Microsoft Excel (Data Analysis -> Descriptive Statistics):

X (Corruption perceptions index)		Y (Software piracy percentage)	
Test	**Values**	**Test**	**Values**
Mean	4,721818182	Mean	59,68181818
Standard Error	0,211245095	Standard Error	2,021029219
Median	4,1	Median	63
Mode	2,2	Mode	25
Standard Deviation	2,215557246	Standard Deviation	21,19673327
Sample Variance	4,908693912	Sample Variance	449,3015013
Kurtosis	2,195560762	Kurtosis	1,098552302
Skewness	0,632666234	Skewness	-0.2784883
Range	7,9	Range	75
Minimum	1,5	Minimum	20
Maximum	9,4	Maximum	95
Sum	519,4	Sum	6565
Count	110	Count	110

From the analysis we can make the following observations:

- Both the corruption perceptions index (x) and the software piracy level (y) have strong variations: from 1.5 to 9.4 and from 20% to 95%
- For the software piracy level (y):
    - The skewness is -0.278; we can calculate τ_1 using the following formula:

$$\tau_1 = \frac{S-0}{\sqrt{\frac{6}{n}}}$$

- Because $|\tau_1|= 1.190<1.96$ we can accept the H_0 hypothesis (the distribution is symmetrical and is accepted at a degree of significance of 5%
- Kurtosis is 1.909, less than 3, so we can calculate τ_2:

$$\tau_2 = \frac{K-3}{\sqrt{\frac{24}{n}}}$$

- Because $\tau_2 =-2.336$ and $\tau_2 <-1.96$, then H_0 is rejected for a degree of significance of 5% meaning that the distribution is platykurtic

* For the corruption perceptions level (x) we have:
  - Skewness = 0.632, $|\tau_1|=2.681$, so $|\tau_1|>1.96$ => H_0 is rejected for a degree of significance of 5%, so the distribution is asymmetrical to the right side
  - Kurtosis = 2.195, $\tau_2= -1.708$ and because $-1.96 < \tau_2 < 1.96$ => platykurtic distribution

We can check to see if **x** is affected by measurement errors, by using the formulas:

$x \in (\bar{X} \pm 3\sigma_x) \Leftrightarrow \bar{X} - 3\sigma_x < x_t < \bar{X} + 3\sigma_x \Leftrightarrow$

$4.72 - 3 * 2,21 < x_t < 4.72 + 3 * 2,21$

$-1.910 < x_t < 11.350$

x_t is replaced in turn by the minimum and maximum values 1.5 and 9.5:

-1.910 < 1.5 < 11.350 (true)
-1.910 < 9.5 < 11.350 (true)
Because both statements are true, we can safely say that **x** is not affected by measurement errors.

We can check to see if **y** is affected by measurement errors, by using the formulas:

$$y \in (\bar{Y} \pm 3\sigma_x) \Leftrightarrow \bar{Y} - 3\sigma_y < y_t < \bar{Y} + 3\sigma_y \Leftrightarrow$$

$$59.68 - 3 * 21,19 < y_t < 59.68 + 3 * 21,19$$

$$-3.890 < y_t < 123.250$$

y_t is replaced in turn by the minimum and maximum values 20 and 95:

-3.890< 20< 123.250
-3.890< 95< 123.250
Because both statements are true, we can safely say that y is not affected by measurement errors.

Using the OLS (least squares) method to estimate the parameters

Results obtained using EViews:

Dependent Variable: Y

Method: Least Squares
Date: 11/19/10 Time: 10:28
Sample: 1 110
Included observations: 110

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	-8.130372	0.485234	-16.75557	0.0000
C	98.07196	2.528793	38.78212	0.0000

R-squared	0.722186	Mean dependent var	59.68182
Adjusted R-squared	0.719613	S.D. dependent var	21.19673
S.E. of regression	11.22400	Akaike info criterion	7.692000
Sum squared resid	13605.64	Schwarz criterion	7.741099
Log likelihood	-421.0600	F-statistic	280.7490
Durbin-Watson stat	2.270432	Prob(F-statistic)	0.000000

Results obtained using Microsoft Excel:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,849815
R Square	0,722186
Adjusted R Square	0,719613
Standard Error	11,224
Observations	110

ANOVA

	df	SS	MS	F	Significance F
Regression	1	35368,23	35368,23	280,749	8,24E-32
Residual	108	13605,64	125,9781		
Total	109	48973,86			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	98,07196	2,528793	38,78212	3,2E-65	93,05945	103,0845	93,05945	103,0845
X Variable	-8,13037	0,485234	-16,7556	8,24E-32	-9,09219	-7,16855	-9,09219	-7,16855

Applying statistical tests:

The regression line

The regression line is $\hat{y} = a + bx$, where b=-8.130372, a=98.07196.

The slope b = -8.130372 suggests that if the corruption perceptions index modifies by 1 point (meaning that the corruption level decreases and the citizens' trust rises) the software piracy percentage lowers by 8.13 percentage points.

The interception point a = 98.07196 is the point in which the regression line intersects the Oy axis, meaning that when x = 0 the values of y is 98.07196. In other words, in a perfectly corrupt country, the software piracy level is almost 100%.

The standard error for the regression

For every value of x, we calculate the value of $\hat{y}$: ($\hat{y}_i = 98.07196 - 8.130372 x_i$) and for every y_i we compute the following difference: $y_i - \hat{y}_i = e_i$. The Sum of Square of Error = $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$.

The residual variables' dispersion is: $s_e = \sqrt{\frac{SSE}{n-2}}$ meaning s_e^2 = 13605.64/(110-2) => s_e =11.22400.

The lowest value that s_e can take is 0, when SSE=0 (all the points are on the regression line). So, the lower s_e is the less far away from the regression line the value is and the better and more accurate the prediction.

Interpreting the value of s_e is done by comparing it to the dependent variable y, more exactly, to the average of the series, $\bar{y}$.

Because s_e =11.22400 and $\bar{y}$ = 59.68182 we must admit that the standard error for the regression is quite large. We cannot evaluate the model based on s_e because there is no upper limit set for s_e.

1. The F statistic

The two hypothesis are:
$H_0 : s^2_{y/x} = s^2e$, meaning the two dispersions are approximately equal, so the influence of the x factor does not differ from the influence of random factors;
$H_1 : s^2_{y/x} \neq s^2e$, meaning that the influence of the x factor and the influence of random factors, measured by the two dispersions, differ significantly;

Testing the significance of the two dispersions is done using the F test. Knowing the two values F_{calc} and $F_{\alpha,v1,v2}$ (which is the theoretical value for the F variable, taken from the Fisher – Snedecor repartition table, at a degree of significance α and a number of freedom degrees v1 = k; v2 = n-k-1), de rule for the decision is: H_0 is accepted and H_1 is rejected if $F_{calc} \leq F_{\alpha,v1,v2}$.

The value for the F statistic is F_{calc} = 280.7490 $F_{(0.05, 1, 153)}$ = 3,9290114, so $F_{calc} > F_{(0.05, 1, 153)}$ and Prob(F-statistic) is very small (0.000000), which means that H_0 is rejected, H_1 is accepted, which means that the regression model is statistically significant, it is valid..

The coefficient of determination

R^2 = 0.722186

This statistic shows that 72.21% of the y variable is explained by the variation of x. The coefficient of determination strengthens the conclusion that there is an obvious linear relationship.

We can get the value for the correlation coefficient from the correlation matrix r_{xy} = 0.9999 which shows that there is a strong positive correlation between x and y.

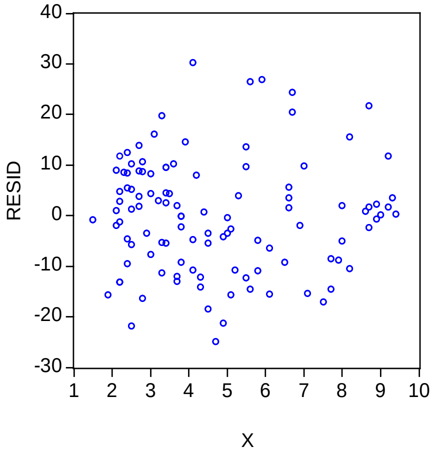
We can get the value for the correlation coefficient from the correlation matrix r_{xy} = -0.849815 which shows that there is a strong negative correlation between x and y.

	x	y
x	1.00000	-0.849815
y	-0.849815	1.00000

2. The Durbin-Watson statistic

The regression model is: y = a+bx, and the following hypothesis are made:
H_0: $\rho = 0$ (the coefficient for the autocorrelation of errors)
H_1: $\rho \neq 0$
The obtained value is d=2.27, so we have D_L =1,65 and D_U =1,69. The following equation must be verified:
$D_U < d < 4 - D_U$
1.69 < d < 4 - 1.69
1.69 < 2.27 < 2.3

$D_U < d < 4 - D_U$ is verified which means that the residues are independent.



Conclusions

We can see, from the regression model and the statistical tests applied, that there is a real correlation between the Corruption Perceptions Index and the Software Piracy Rate of any given country. This means that the assumptions made at the beginning are true and no matter how well protected the software product is, it is still going to be illegally distributed. Maybe this is the conclusion that most software companies arrived to also and this is what made even the most unlikely ones to turn to open software principles. Maybe the overall problem is not piracy, but the fact that intellectual property is not something that can be or should be imposed, but it is a good for all humanity. Money can be made from other sources than selling collective intelligence and many companies are starting to see this. We can only hope that music companies and film makers are going to realize this also in the future.

Bibliography

1. Andrei, T., Bourbonnais, R., Econometrie, Editura Economică, Bucureşti, 2008
2. Andrei, T., Spircu, L., Aplicatii in econometrie, Editura Economica, Bucuresti, 2009
3. Andrei, T., Stancu, S., Iacob, A., Tuşa, E., Introducere În Econometrie Utilizând *Eviews*, Editura Economică, Bucureşti, 2008
4. http://www.transparency.org
5. http://portal.bsa.org
6. http://en.wikipedia.org/wiki/Software_piracy
7. http://en.wikipedia.org/wiki/Corruption_index
8. http://en.wikipedia.org/wiki/Open_source
9. http://en.wikipedia.org/wiki/Free_software