INAPPLICABILITY OF THE ORDINARY LEAST SQUARE METHOD FOR THE WEB. THEORETICAL CONCEPTS

CARAGANCIU Iulian[1]

Lucian Blaga University of Sibiu

Abstract:

This paper aims at presenting the Ordinary Least Square regression analysis used by Antitrust Authorities, in order to determine the dominant position of a company or a newly formed merger. We try to present the inapplicability of such for the Web market due to factors such as uncertainty while determining the dependent and independent variables.

Keywords: *Ordinary least square method; OLS; Web Market.*

JEL classification: *C15; C01; C92*

Introduction

The ordinary least square regression is a method frequently used by antitrust authorities in order to determine whether the company, which is being researched, is able to practice an increase in price and still not lose market share or lose a small and insignificant amount of the latter. The same method is also used in order to analyze whether a merger could pose a threat to the market competition and distort it. The method functions on the premise that we have two variables which need to be tested, whether one influences the other or not.

[1] PhD Student, Department of Economics, Faculty of Economic Science, University Lucian Blaga, Sibiu, Romania, e-mail: j.caraganciu@gmail.com

This method has its flaws and difficulties. One of these that is mentioned in this paper is posed by the fact that Antitrust authorities do not always have the necessary data from the companies for full analysis. On the web to make such an analysis becomes even harder and close to nigh impossible.

At the end of this paper we present a way a two stage game could be structured in order to model the web, and then the results of this game could be tested by regression analysis as to show the relevance and the importance of the factors involved in this game for the web market.

The principle of OLS regression (Ordinary Least Squares)

Multiple regression analysis can be used as a useful statistical tool that can quantify the effects of multiple variables on the outcomes of interest. Unfortunately economists in competition authorities cannot run experiments on the field (Stillman, R. (1983)). Therefore they cannot dictate real companies to increase their prices by 5% and analyze the results of the increase in price on the real market. However, sometimes such data is present in the past of a company, and can prove an invaluable tool for market analysis since it gives us the real company and competitors' decisions and behavior. In this case while we can see real reactions from competitors and the behavior of the market, the researcher wouldn't have that much of control over the variables going on the market, this therefore could complicate the research (Stone, M., & Brooks, R. J. (1990)).

1.1 Data generating process and Regression Specifications

The start of a regression analysis is generated by the assumption that the analyzed variables have a relationship between them. I.E. we assume that price has an effect on quantity demanded of a certain good. The quantity demanded by the market of a certain good will be denoted *Q* and the price will be denoted *P* then the relationship between these two variables will be given by the expression:

$P_i = a_0 + b_0 Q_i + u_i,$

Where I indicates the different possible observations of reality at any given moment I and the parameters a_0 and b_0 take on real values. This presents the DGP (data generating process) where u_i presents the shock not known to the analyst. Regression analysis is based on the premise, that given enough observations of P and Q we can learn about the true parameters a_0 and b_0 without observing the u_i.

These basics where necessary in order to better understand the assumptions and limitations, as well as conditions, of the Ordinary Least Square method. The data generated in the way described above presents real potential for analysis since it has to be made from observations from the real market and therefore might be more relevant than lab generated data.

1.2 Ordinary Least Squares Method

Consider the following regression model:

(1) $y_i = a + bx_i + e_i$

The OLS regression estimator attempts to estimate the effect of the variable x on every var y by selecting the parameters (a,b). In order for the analysis to take place the method of OLS assigns maximum possible explanatory power to the variables specified as determinants of the outcome. This is done in order to minimize the effect of the residual component e_i. The value of the residual component depends on the choice of variables established so therefore can be written as $e_i(a,b) = y_i - a - bx_i$. Formally OLS will choose the parameters a and b as to minimize the sum of squared errors, therefore it can be written as follows:

(2)$\min_{a,b} \sum_{i=1}^{n} e_i(a,b)^2$.

The method of least squares is quite general. The model described is linear, but the technique can be applied more generally. For example we may have a model which is not linear in parameters, which states $e_i(a,b) = y_i - f(x_i;a,b)$, where for example $f(x_i;a,b) = ax^b$. The same least square method can be used to estimate the parameters.

If the model is linear that we have the Ordinary Least Squares method while if the model is nonlinear then the method is called Nonlinear Least Squares (NLLS).

The linear model assumes that we have linear parameters and linear variables in a manner that a change in the variable x will always produce the same amount of change in the variable y. I.E. this can be that the variable x to express the price P_i and the variable y to express the quantity Q_i. Then if the price for a good is increased by 0.50 $ then the reduction in quantity Q_i will be four units of good per month. Given the linear parameters and variables assumption this relationship will hold whether the initial price was 5.00$ or 7.00$, which is not always correct when modeling the real market.

In order to solve this problem it is common to operate a log transformation on price and quantity variables so that the decrease or increase

in variable y will be measured in percentages. Therefore $y_i = lnQ_i$ and $x_i = lnP_i$. In this case we will have $\frac{\partial lnQ_i}{\partial lnP_i} = b$while $\frac{\partial Q_i}{\partial P_i} = \frac{bQ_i}{P_i}$, therefore the absolute changes will depend on both levels of price and quantity demanded. These variable transformations do not change the fact that the parameters and variables are linear and therefore will still be solved using the OLS method.

Let $(\hat{a}, \hat{b})$ be estimates of the parameters a and b. Then the predicted value for y_i given the estimates of a and b and a fixed value for x_i will be as follows:(Peter Davis and Elena Garcias 2010)

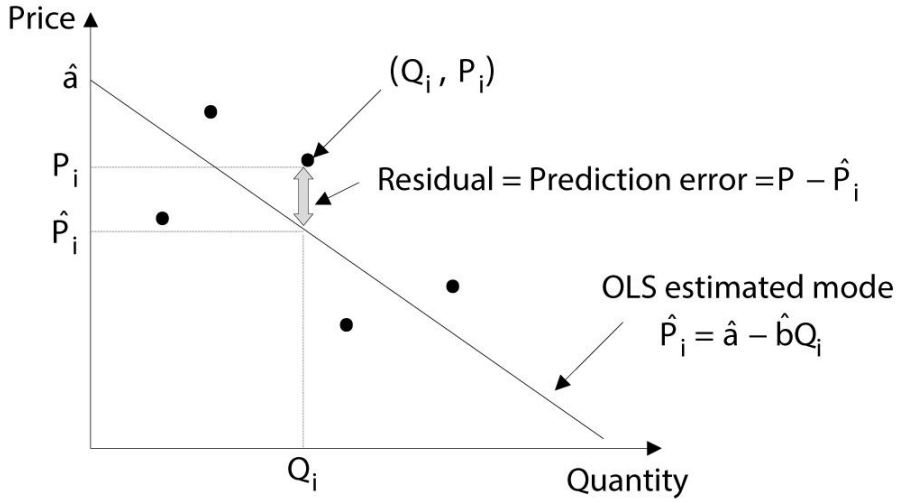
(3) $\hat{y_i} = \hat{a} + \hat{b}x_i$.

The difference between the estimate $\hat{y_i}$ and the actual y_i is the estimated error e_i. Thus we can write this as:

$e_i = \hat{y_i} - y_i$.

Figure 4.4 shows us the estimated residuals for the inverse demand curve, where y_i is the quantity Q_i and x_i is the price P_i.

Figure 4.4 Estimated Residuals in OLS regression



Source: Source: Peter Davis and Elena Garcias 2010 – "Quantitative Techniques for Competition and Antitrust Analysis" – Princeton University Press, Princeton and Oxford p 68

The positive residuals are above the estimated line and the negative residuals are below it. The estimation of the inverse demand function, in the

OLS method, minimizes the total amount of squares of vertical prediction errors. In case the model uses a true Data Generating Process and the parameter estimation is right then the residuals will be exactly as the real errors (for instance shocks in our case).

Mathematically finding the OLS estimators requires solving the minimization problem:

(4) $\min_{a,b} \sum_{i=1}^{n} e_i(a,b)^2 = \min_{a,b} \sum_{i=1}^{n} (y_i + a + bx_i)^2$.

The first-order conditions are given by setting the first derivates with respect to a and b to 0:

(5) $\sum_{i=1}^{n} 2\left(y_i + \hat{a} + \hat{b}x_i\right)(-1) = 0 \; and \; \sum_{i=1}^{n} 2\left(y_i + \hat{a} + \hat{b}x_i\right)(-x_i) = 0$.

The minimization problem is quadratic in the parameters, and hence the first order conditions are linear in the parameters. As a result the first order conditions for linear parameters provide us with a system of equations to solve, the number of equations is equal to the number of parameters.

If we have 2 parameters the first equation can be written as $\hat{a} = \bar{y} - \hat{b}\bar{x}$, where $\bar{y}$ and $\bar{x}$ denote sample averages as we are to show further:

(6) $\sum_{i=1}^{n} 2\left(y_i + \hat{a} + \hat{b}x_i\right)(-1) = 0$

(7) $\Leftrightarrow \sum_{i=1}^{n} y = \hat{a}n + \hat{b}\sum_{i=1}^{n} x_i$

$$\Leftrightarrow \hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i.$$

In case we get the $\hat{b}$ parameter equal to 0 then the parameter $\hat{a}$ is represented by the average value of the dependent variable.

Sometimes we have the case where the dependable variable is explained by a number of explanatory variables. I.E. price can be explained by cost and by brand image at the same time but in different manner. This leads us the conclusion that the secondary variable can be of low order or higher order than the first variable, meaning that this can account for both multiple variables and particular nonlinearities in variables. Retaining the linear-in-parameters specification, a multivariate regression takes the form as follows:

(8) $y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + e_i$.

For given parameter values, the predicted value of y_i for given estimates and values of (x_{1i}, x_{2i}, x_{3i}) is:

(9) $\widehat{y_i} = \hat{a} + \widehat{b_1} x_{1i} + \widehat{b_2} x_{2i} + \widehat{b_3} x_{3i}$

And the prediction error is $e_i = y_i - \widehat{y_i}$.

The minimization problem in this situation is the same as before, the only difference being that we have more variables to minimize over. The minimization function takes the form of:

(10) $\min_{a,b_1,b_2,b_3} \sum_{i=1}^{n} e_i(a, b_1, b_2, b_3)^2$.

Provided the model is linear in parameters, the minimization function will take a quadratic function just like in the case of a single variable optimization, and the first-order conditions will also be linear, which will allow for analytic solutions.

In order to determine these solutions it is easier to use the matrix notation, according to the unifying treatment provided by Anderson (1958). This will give us a matrix of the form:

(11) $$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \beta + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

Which in turn can be more easily expressed in terms of vectors and matrices as:

(12) $y = X\beta + e,$

Where y is an$(n \times 1)$ vector and X is $(n \times k)$ matrix of data, while β is an $(k \times 1)$ vector of parameters to be estimated and e is a vector of residuals. For our example we have k=4 since we have four parameters to be estimated.

2 Common errors in the OLS method

Here we are going to treat the most common errors that occur during a regression analysis. More precisely we are going to discuss: misspecification, endogeneity, multicollinearity, measurement error and heteroskedasticity.

a) *Misspecification* generally it occurs when the regression model cannot represent the true Data Generating Process. In other words the representation of the world, by the model is not valid to the economic environment it is supposed to represent.This might happen due to the analyst, while specifying the regression model, creates imposed restrictions between the variables that do not hold true. This kind of error occurs sometimes, due to the incorrect functional form in the relation between two variables. I.E. we might accidentally specify x instead of $\ln x$.

Another source for this error to occur is the possible omission of an important explanatory variable, thus setting its value to 0 in the regression. This leads to making the entire model flawed. I.E. a term with a higher order could have been omitted.

Basically the misspecification error is related as the name suggests with misspecification of the parameters or the connection or the order of connection between the later. This is an error that might occur due to a market that is hard to model and understand completely or due to a market having a specific that is not covered by the model.

b) *Endogeneity*means that one of the regressors used in the model is correlated to the shock component of the same model. This might lead to serious errors in the model's outputs and interpretation. I.E. an included regressor might be irrelevant but still correlated to the causal factor, which has been omitted from the regression.

Another case when endogeneity occurs is when the regressor and the explained variable are determined simultaneously. In other words this means that the regressor and the explained variable change when the shock occurs (I.E. Demand estimation).

In such a case the solution could be to explicitly model the full system of equations rather than taking into account just a single estimation gained from a single equation. I.E. in demand estimation context the pricing equation might be added.

Models in which the full system of equations is taken into account and are full and explicit models and where all of the endogenous variables are known are called "full information" models. On the other hand when we estimate just s ingle equation instead of the whole system it is known as "limited information" estimation since they do not always provide a full and complete picture.

c) *Multicollinearity* occurs when the explanatory variables in a regression are highly correlated with each other, therefore the effects of different regressors will not be separable and the estimators will not represent the true effect of the variable on the outcome. Assuming a true DGP is:

(13) $y_i = a_0 + b_{10}x_{i1} + b_{20}x_{i2} + u_i.$

If $x_{i1} = \lambda x_{i2}$, then the DGP can be rewritten as

(14) $y_i = a_0 + (b_{10} + \lambda b_{20})x_{i2} + u_i.$

The main problem with this specification is that we cannot define the separate effects of x_{1i} and x_{2i} on y_i. We can only identify the combination $b_{10}\lambda + b_{20}$.

More precisely the two variables do not necessarily have to be linear combination of each other but it is enough if they are rather close. This could pose a problem if the sample is small.

d) *Correlated errors and Heteroskedasticity*

Correlated errors presume the correlation of error terms across observations can arise in a number of contexts. The most eloquent case being with time series, when shocks take several periods to fade, therefore a negative shock in the first period would lead to a negative shock in the next etc. In order to reduce the effect of this error, judging by the OLS formula, we could duplicate the information, but clearly, even though the effect of the error would be reduced, we would not get new data, it will simply be the duplicate of the existing one. The reason this could lead to larger problems is that the OLS estimates of standard errors relies on the assumption that all of the data represents independent realizations of the variables, whereas in this case the fact that the data is duplicated will lead to a very strong dependence.

The assumption of heteroskedasticity is the assumption that $E[u_i^2] = \vartheta_u^2$ for all i. When this assumption fails we will be described as being in a situation of heteroskedasticity where $E[u_i^2] \neq E[u_j^2]$ for some $i \neq j$. In other words the situation of heteroskedasticity is when the variance of the error is different throughout observations. On the other hand the heteroskedasticity of errors can be when different groups of observations, which might have different variance of the errors. As for the OLS method, when the variance of the error terms is not homoscedastic, the standard errors calculated by the OLS will usually be wrong. Unfortunately heteroskedasticity is a potential problem in the OLS method unless the correct formulas are used, therefore the analyst has to be careful as to do so.

The OLS method presented before does not prove a very useful tool when analyzing the web market, since the observations on this market are not consistent with price, but rather with the willingness of the consumers to buy online, fact that can be seen by sheer sale numbers. The main problem with the adaptation of the OLS method to the web market could be posed by the fact that it is unclear what should be correlated with what, since the sales aren't heavy price dependent. If they were then the market would have a larger

share than that of the real market due to the lower prices and the commodity of receiving the good without leaving your home.

3 Two Stage Web Market Game introduction

The regression method is the least suitable tool for analyzing the web market, since the correlation has to be clear and only tested. While on the web market the main problem is to find the variables that have to be correlated. The correlation between the amount of sales online and the willingness to buy online is pretty much evident. The main point here is that the web market is a more behavioral driven market.

The Bertrand price based model can be applied in order to model a stage of the web market and the Cournot quantity equilibrium could be implemented for this market, the OLS method could prove a useful tool for hypothesis analysis, since it could show whether a factor A has an impact on the factor B.

This entire material was presented in order to further develop a two stage game which would show the way the web market works and potentially reach equilibrium on the market.

Since on the web market even though the prices might be equal and the output quantity of the companies might be the same, the market still isn't divided among competitors in an equal share. This is due to the market competitive structure presented before, which involves that the company first of all has to reach a level of awareness and then participate in the balancing levels of trust among companies. After the awareness of a company as well as the level of trust for a web company are the same, then the reputation comes into play. The level of reputation of a certain company is strongly related to the level of price that web-company can practice, due to the fact that the web market is a differentiated-service based market.

Therefore the higher the level of reputation of a web company the higher the prices it can practice without losing market share.

The goal of the two stage game would be to simulate the way the companies on the web set their prices according to the level of awareness, trust and reputation they gain. The type of game however is going to be a lab experiment with afferent observations and conclusions based on those observations.

The control groups will be fed different information on the companies, in order to simulate the level of awareness of the virtual companies as well as

the levels of trust and reputation. This will help show consumer preferences related to the awareness levels of the companies. Another idea would be to show the role of social networks in the increase of the awareness level of a certain company as well as for gaining a general idea of the speed of information transmission about a certain web company.

The awareness can be modeled by feeding just a part of the participants with information about, let's say companies A and B, and telling each participant about the company C, which in turn would increase the awareness of company C among the estimated consumers.

The modeling of the level of trust can be achieved via, giving out information about the background of the company as well as about the previous orders that were fulfilled by this company. The reputation modeling would prove to be the trickiest as it would have to be fed by the quality of the services as well as by the time the company is present on the market.

After the first stage is finished and the control groups are given the information, we would then begin the second round, by giving the participants a fixed amount of virtual money and then ask them to buy a certain product from one of the companies on the web market and feed them flawed information that one of the companies from the market is a scam and will receive the payment but will not deliver the good.

After this we could get a general idea on the importance of trust and reputation as well as awareness on the web market.

Another option would be to make the players participate in repeated purchases from the market with one of the companies not delivering the good, and some of them delivering the service at a particularly high level. Observing this would give us information on the increases and decreases of trust, reputation and even awareness on the market, as a general idea.

Conclusions

In conclusion we could state that while OLS is not directly applicable for the web market it could still prove a useful tool for analyzing correlation between factors presented in web models and games. This could prove a powerful testing mechanism for determining the relevance of variables in a model, lets say on market price or on the amount of customers.

In this paper we also presented some theoretical background on how a two stage game could be made as to model the web medium. This could prove to be a beginning for future research in the matter.

References:

- Stone, M., & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 237-269.
- Peter Davis and Elena Garcias 2010 – "Quantitative Techniques for Competition and Antitrust Analysis" – Princeton University Press, Princeton and Oxford
- T.W. Anderson – An introduction to multivariate statistical analysis, 1958
- Stillman, R. (1983). Examining antitrust policy towards horizontal mergers.*Journal of Financial Economics*, *11*(1), 225-240.