UDC: 510.647
Professional paper

ANTICIPATION OF DYNAMICS AND IDENTIFICATION OF POTENTIAL FACTOR VARIABLES A CASE STUDY

Elmira KUSHTA [1], Miftar RAMOSACAJ[1]

[1]*Department Mathematics, Faculty of Technical Sciences, University of Vlora*
**Corresponding Author: e-mail: elmira.kushta@univlora.edu.al*

Abstract

Based on this paper we have identified the true factor variables [X] and that response [Y] for the given system the first variable does not give a good match for any combination. This response variable is inappropriate for judging the system as a whole. Taking another variable that resulted in a good logic regression, we see that the elimination of the first 2 variables does not significantly change the fit of the regression. Again, the regression is not very good, which indicates the presence of other factors included in this study, but for the case with 5 variables, the pepper has a good deal with all the variables We note that the technique used with linear OLS and non-linear OLS can be well adapted to the logistic case. By imposing a practical balance between regression fitting goodness and the quality of the reproduction of the values for response
variables, we decided to set-up a well-designed logistic model to analyze consumer behavior for a real market (district of Vlora). The steps proposed herein have worked as a data–oriented modeling methodology that reduces subjective or empirical approaches in econometric and marketing analysis.

Keywords: variable, consumer behavior, logistic model, linear OLS

Introduction

In this work, we elaborate on some ideas to simplify the analysis of consumer behavior in small markets supposedly under the effects of heterogeneities and undersized measurement data. Herein we used the factorial analysis to reduce the initial set of variables to a more compact and representative subset, and among them, we next selected mainly stationary variables to proceed with calculations. By imposing a practical balance between regression fitting goodness and the quality of the reproduction of the values for response variables, we decided to set up a well-designed logistic model to analyze consumer behavior for a real market (district of Vlora). The steps proposed herein have worked as a data–oriented modeling methodology that reduces subjective or empirical approaches in econometric and marketing analysis. So, by a combination of very simple descriptive statistics tools, the complexity of the behavior of the consumers and thereafter its modeling dimensions were finally both reduced and the analysis was improved. The study of concrete economic environments unavoidably faces challenges for scholars. Standard questionnaires that aim to gather information from social or economic mediums include different types of variables, non-numerical responses, questionable answers, missing or incomplete records, etc. Usually, the inquiries might be organized and held

at different moments in time. Finally, they must be included in modeling say linear multivariate functions

$$Y = A * X + \varepsilon; \quad (a)$$
$$Y = W * Z + u \equiv W * (\Gamma * X + v) + u \quad (b) \qquad (1)$$

or in the logistic type relationship

$$f(z) = \frac{1}{1 + \exp\left(-\alpha + \sum_{i=1}^{n} \beta_i x_i\right)} \qquad (2)$$

where X, are factor variables or predictors and Y are response variable or indicators whereas u,v,ε are errors and A,W, Γ etc are matrices. Versions 1. (a) are the simplest relationships in the models. In case 1.1 a regression procedure leads to the calculation of the matrices of coefficients which explain the weight of each variable (i) in responses (j). In the case of 1.1. (b) the problem includes the calculation of the so-called latent variable (Z) adding to the coefficient's matrices W and Γ. The belongs to the structural equation or SEM systems, discussed in [6], [7] and used largely in sociology [10], econometrics [11], explanatory medicine, etc. In all those cases, some necessary statistical assumptions should be fulfilled. It happens that in real systems many of them do not hold. Therefore, quantitative methods need for more analysis as seen in the reference [9] and others related to these aspects. In this case, some approximate methods are suggested and elaborated as in [10] or in a more dedicated case in time series in [12]. However, in general, it depends on the concrete properties of the data series. In general, preparatory analysis or data elaboration is needed. The second problem is related to the tangible set of the variables included in the models. Again, standard models belong to the standard systems, and in real ones, there is a considerable difference. But by carefully using simple analytic tools it is possible to avoid the complexity of the model, to control extra errors added during the calculation phase, and to improve overall calculation. In our recent research in the analysis of consumer behavior in the district of Vlora, we considered such specifics as an important step [1], [2], etc. The last issue that can affect directly the quality of the modeling is the representable property of the data gathered from measurement related to the sampling process. In practice, an appropriate size of the sample might not be accessible [3] or it is difficult to be stated. For numerical continuous and normally distributed random variables, the working formula is

$$n = Z_{\frac{\alpha}{2}}^2 \frac{p(1-p)}{M \arg in\_of\_Error} = \frac{\frac{z^2 p(1-p)}{e^2}}{1 + \frac{z^2 p(1-p)}{e^2 N}} \qquad (3)$$

where z is the normalized variable $z = \frac{x - \langle x \rangle}{\sigma(\langle x \rangle)}$, Z is the critical value or level α, N is the population size and e are the level of tolerance adapted and p is the sample proportion. However, it is difficult to estimate if we consider categorical variables. In this case, we proposed to choose a sample size according to numerical variables and accordingly to use auxiliary statistical tools to identify the error injected in the system by such an approach. Next consider that in theoretical approaches one assumes stationary for the system states, homogeneity, formal relationship, etc. Detailed analysis on those aspects are provided in many articles-guides and statistical books as [3] or [4]. In this case, the problems could be overcome if we adjusted correctly the sample size or adopted a suitable sampling method. In the case where the above step is not suitable or even

impossible, the factorial and descriptive analyses could be used as recommended in standard procedures to manage the sampling error, [5] and general consideration [13].

1. Statistical Models

Different statistical models use different formulas to find the study weights. The main focus of every statistical model is to re-distribute the weights appropriately in the meta-analysis. The main objective of the re-distribution of weights is to find the most precise estimate.
If the studies are not heterogeneous the fixed effect model is fine. Unfortunately, in real life most of the studies are heterogeneous, and hence fixed effect model is not appropriate. As part of the process test of heterogeneity is essential. This is done by using Cochrane's Q statistics which follows a Chi-squared distribution. If the test outcome is significant the fixed effect model is inappropriate and requires an appropriate statistical model.
Every model assumes that there is a common effect across all studies, and pulls data from all studies to estimate the unknown common effect size, usually denoted by Θ using an appropriate statistical model.

1.1 Relaxation of the distribution after sales: We start by the inspecting statistical behavior of the system by analyzing the distribution for characteristic variables, and we observe that among many candidates the q-Gaussian mentioned before in our work and introduced by [13],[14], etc

$$p(x) = \alpha\left[\left(1-\beta(1-q)(-\mu)\right)^2\right]^{\frac{1}{1-q}} \qquad (4)$$

fits the data better than other tested. In addition and to count for mixed multiplicative properties as usually expected for complex dynamics, we use q-lognormal as detailed theoretically in the reference [14]

$$p(x) = \alpha\frac{1}{x^q}\left[\left(1-\beta(1-q)\left(\left(\frac{x^{1-q}-1}{1-q}\right)^{1-q}-\mu\right)\right)^2\right]^{\frac{1}{1-q}} \qquad (5)$$

Q-distribution has been successfully used in and suggested in the studies for complex systems as generalized in the reference [14] and applied in [15] etc. Remember that q-additive and q-multiplicative processes responsible for q-Gaussian and q-lognormal respectively, are defined by q-algebra as follows

$$a \oplus_q b = \begin{cases} a+b+(1-q)ab & a,b>0 \\ 0 & a,b\le 0 \end{cases}$$
$$a \otimes_q b = \begin{cases} [a^{1-q}+b^{1-q}-1]^{\frac{1}{1-q}} & a,b>0 \\ 0 & a,b<=0 \end{cases} \qquad (6)$$

that clearly shows the q parameter is the measure of the distance from pure processes. If in (5) we denote a,b,c.. the probability for separate events, the probability of their occurrence becomes
quite complicated as happens in reality. If a process has two types of interactions say the additive and multiplicative properties, the distribution characterizing the state is more likely to be a q-Gaussian [16]. A q-lognormal will be considered if the logarithm of variable is considered but the original equation in 5 does not

show directly the expected q-multiplicative behavior. Moreover, the q-lognormal is a stable attractor only for q=1, therefore functions of type (3) are very sensitive toward the q-parameter. For a correct use of such analysis, we implement a careful bin optimization as described in [17]. After this short briefing with q-distribution applied, we expect to identify the characteristic observables for the responses of consumers in the system.

Here we selected for preliminary analysis the number of consumers visits in the market and the average purchasing expenses. The number of visits is very important for the statistics because it reports the overall reaction of the consumers without being limited form the budget constraints. The distribution of the expenses is by nature the most important econometric parameter. The average and the variance for parameters are measurable if the distribution is stationary so we can perform statistical analysis in stationary states or in those states where variance is finite. According to [13] this would happen when the q-Gaussian parameter is q>5/3. In [14] the full alternative approach called q-Central Limit Theorem stated rigorist of next, we would like to know the attracting property of marketing activity which can be measured by the number of visits; therefore both cases have been analyzed from the stability point of view.

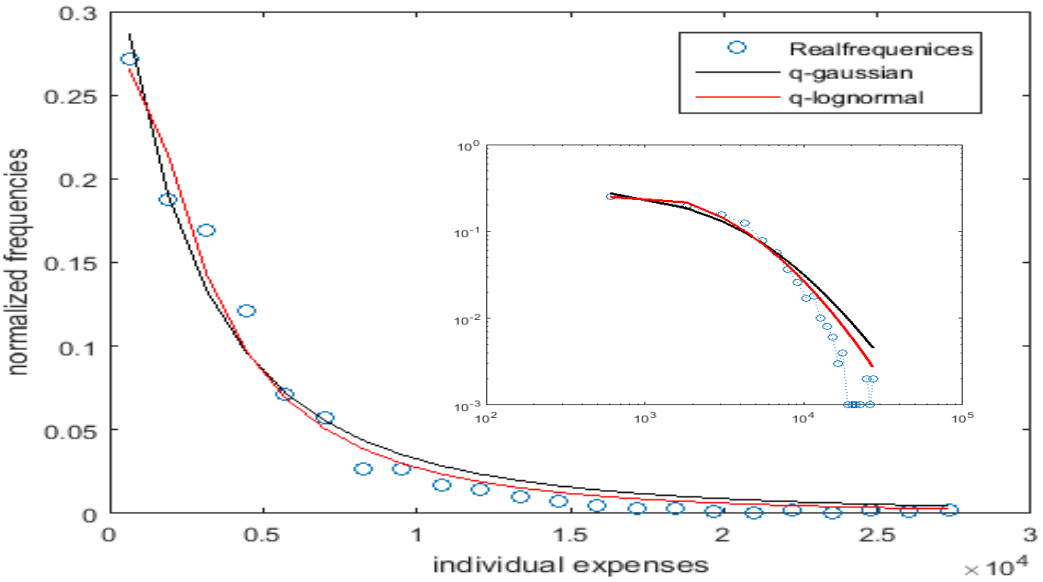

Figure 15: Distribution of visits average expenses.

The small picture shows a log-log representation for a better picture of the fit. After realizing the fit to some expected common distributions, we observe that the parametric q-distribution had the best statistics of the fitting. The fitted curves are mostly q-lognormal within alpha=0.05 restriction, whereas q-Gaussian has lower statistics, but is much less sensitive to the binning assize. Accepting functions of type (2) as best-fitted distributions, one can admit that the processes underlying the expenditures dynamics are q-multiplicative and, hence very complicated. From the fitted q-lognormal we obtained the parameter q=1.0001 which reports a nearly stationary lognormal if multiplicative processes are determinant. In particular, it does not allow measuring the level of non-stationary as the difference q-1=0.0001 is too small. But in the first equation of (3) we see that q-addition involves additive and multiplicative properties, so for mixed processes it seems to be

more significant. For this reason, we prefer q-Gaussian for the analysis of such behavior. Parameters q and adjusted R-squared are [1.6531 0.9661] for q-Gaussian and [1.0001 0.9742] for the q-lognormal fitted. Therefore q-Gaussian tells that q~5/3 is in the boundary of definition for variances

$$\sigma_q = \frac{1}{(5-3q)\beta} \qquad (7)$$

Next, we considered the data for market visits and average expenditures after sales were applied. We obtain that the expenditure distribution was found in a more stable state. The statistics for q-distributions fitted to the frequencies of consumer visits at the market again support the q-lognormal as best fitted function, but again by changing bin size we observe that q parameter in q-Gaussian changed only slowly whereas for q-lognormal it jumps from the value 1 with high margin. Therefore, we consider q-Gaussians for further analysis. Q-Parameters estimated and R2 for this case are found [1.6525 0.9778] for q-Gaussian and [1.0000 0.9974] for q-lognormal. We see that the stationary parameter q is nearly the same for the two series (before and after sales) but as we explained above the observation time for the second is much lower. So we accept that the state after discounts is more stable.

2. Reduced buyer profile with OLS linear regression.

We first considered the situation with individuals with one dimension in the sense of variables, i.e. characterized by only one variable. In the estimation with OLS, we took into account the effects of the edges of the distributions as discussed in the procedure proposed by us [18]. As a rule, this optimization highlights the linear nature of the connection but does not produce well the behavior of the edges which, being symbolic can be found by maximum likelihood. We note that the technique used with linear OLS and non-linear OLS can be well adapted to the logistic case and compete if we include the weights (W-OLS) in the game. This alternative is especially appropriate in the case of the first consideration of the answer as a numerical variable. Here this calculation is preliminary since the logistic regression as the most complicated one may not show any exact relationship especially due to errors in the conditions of the small sample.

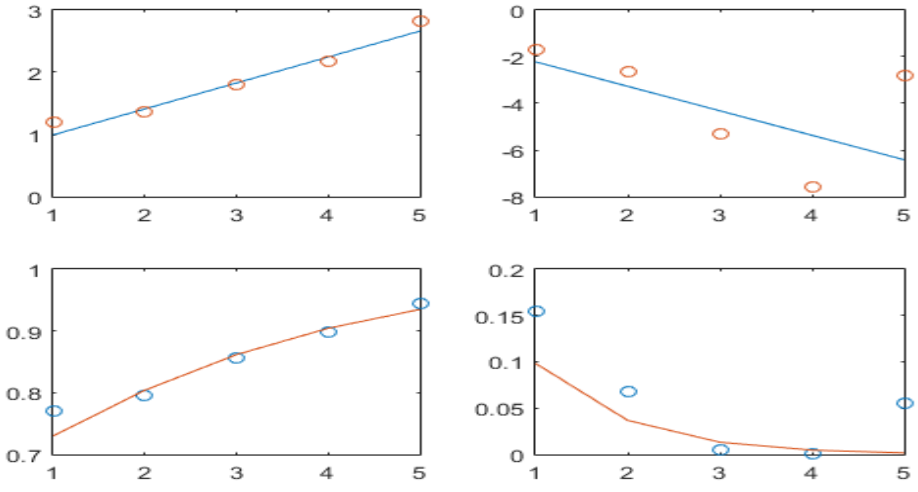

Figure 2: Probabilities and logistic regression for the family category.

The four candidate variables for examining the behavioral system are approximated in the logistic regression as shown in the figure. Referring to the statistical table we see that in all cases the p-value representing the confidence level gives us a result. Statistically, the acceptance that the parameter is constant is overturned since the match must be protected from regression up to the 0.1 level as we discussed above by not using the standard 0.05 level.

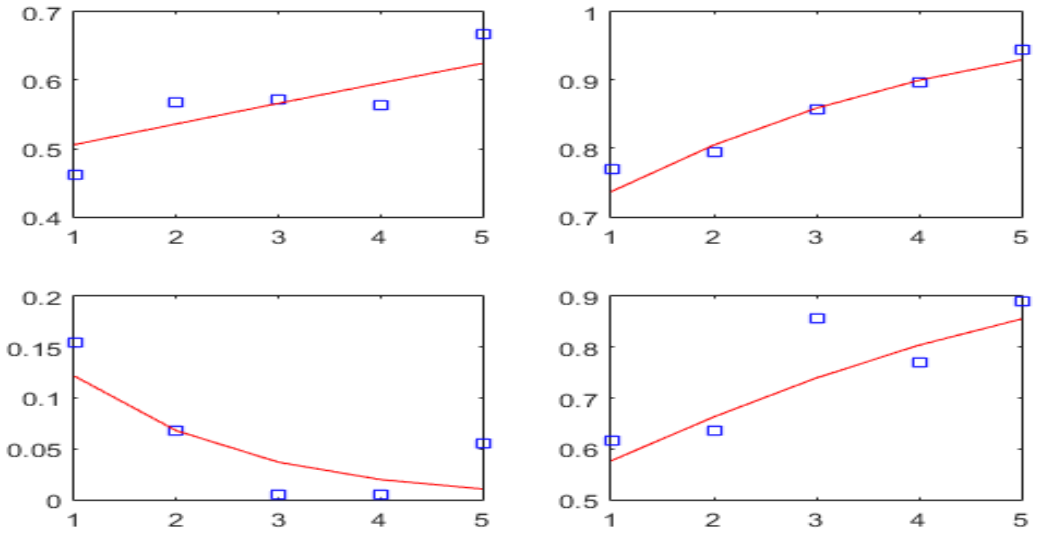

Figure 3: Probability according to logistic regression for the four variables.

We asked for variables to be appropriate for modeling in logistics, MIMIC, and another form if they were found in a stationary state. We managed the measurement realized in the sample where an individual appears as a list of records of different types and different meanings. To include all of them in a deterministic model we must unify their measurement method. Hence categorical variables were transformed using the z-score method in continuous variables. $x \to \frac{x - \langle x \rangle}{\sigma(x)}$. In another step we produced new variable binary by using levels of expenses.

$$ExpencesLevel \leftrightarrow R_i \equiv \frac{Expence\_i}{Total\_Expences} : Y(R_i) = \begin{cases} 1, R_i > 0.5 \\ 0, R_i < 0.5 \end{cases} \quad (8)$$

This last is suitable for logistic and probate modeling

2.1. Factor analysis used in fixing model variabiles: We applied factorial and confirmatory analysis in building logistic and probate models. To define the correct number of variables we performed PCA analyses and identified the number of variables that should be invoked in the modeling logistic or probate. By estimating the variance explained we were able to identify the size of the most important set of variables. It showed that 5 variables could explain more than 95% of the variance. It follows that:

- The system of initially 12 variables resulted reductible.

- Usually 4 groups of expenses are enough to describe the Costumer Behavior

$$ConsumerBehaviour \leftrightarrow \begin{pmatrix} BasicExpences \\ Ordinary\_Substis\tan e\_Expences \\ Life-Qualtiy\_Expences \\ Luxuary\_Expences \end{pmatrix}$$

In table 1 we show parameter of linear equation *LV~ Parameters*Factors*

Table 1: Parameters of hidden variables

	Parameter Matrix A				Parmeter Matrix B	
Factors	Two HV		One H.V	Observed		1 HV-Mode
Free parameter	0.0046	0.526	**0.0046**	SpendingAfterSales		-7.9368
Gender	1.0743	1.3994	**1.0743**	VisitsAftersales		141.482
AgeGroup	-0.2782	83.5887	**-0.2782**			
AverageVisits	-1.8217	-3.5924	**-1.8217**			
Average Spending	-0.3789	-0.4879	**-0.3789**			
RegularClient	0.0363	0.1666	**0.0363**			
TelephoneContact	141.482	0.0798	**141.482**			

Interestingly, the factors have different effects on average spending after sales and average visits. As seen in Table 1, the parameters remain unchanged (up to 3 digits) in modeling with one and two hidden variables and therefore we restrict the model to one single hidden or latent variable. In this case, the hidden variable could act as an interconnection between causes and outcomes. The reduction in the number of latent variables is plausible for the model because in this case, we can use the utility as the intermediate variable or stage in consumer decision-making. So far, we assume that the overall decision of the consumer to increase the expenses after discounts have been applied could be interpreted by a continuous utility function

$$u_j = \beta_o + \sum_{i=1}^{n} \beta_i x_{i,j} \qquad (9)$$

where j is the individual observation and i are variables. The response will be a dichotomous as follows

$$P(Y = y_i | X) = P(a_{i-1} \le u < a_i | X) \qquad (10)$$

and for our binary output, there is only one point to be considered say the moment where the continuous probability takes the value 0.5.

Firstly, we consider the attractiveness of the discounts, so we examine the increasing number of market visits after discounts were applied. Here we use as dependent variables the positive change in the average number of visits after the discount; and for independent factors the gender of the buyer, the age group, and contacts by

calls to announce the offers. By applying probate regression, we observe that a good fit is obtained and the marginal errors are normally distributed as seen in the figure (4). The coefficients have been confirmed as different from zero within 90% confidence, whereas the free coefficient seems to not pass the test.

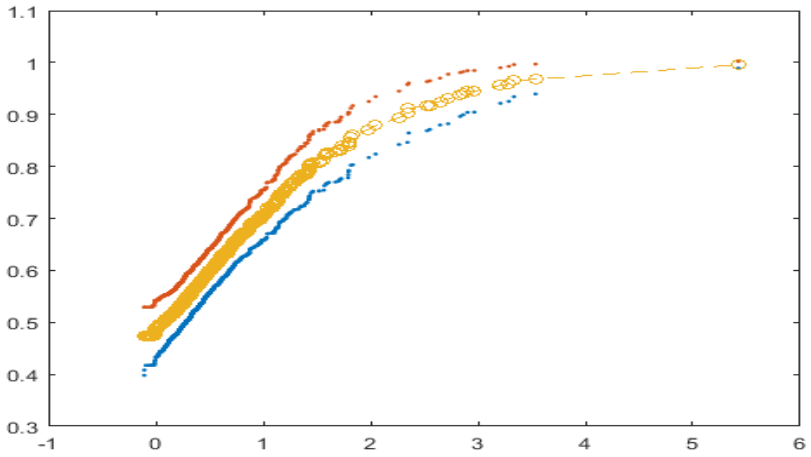

Figure 4: Probate regression for Increasing Expenditures after discounts

Therefore the utility of the attractiveness is obtained by probate regression as follows

$$Y^* = \{0.0698\} - 0.1367 * \text{GenderBuyer} + 0.0962 * \text{Age Group} + 0.0002 * \text{PhoneContact} + \varepsilon$$

From the relation, we observe that the gender of buyers (F=1, M=2) is mostly decisive in the increasing number of visits to the market after sales, and usually, male buyers are not more frequent in the market after discounts have been applied. The phone contact has a slight effect on it. The age group has a comparable role to the gender of consumers. In Figure 4 is seen that the probability of more visits to the market is high for almost all the values of the utility function and only a few values are less than 0.5. In this sense, for nearly all consumers' specifics, the marketing strategy (price discounts) has been found attractive for people who respond by increasing the number of visits to the market. This is the intermediate change in the consumer behavior. In the second stage, the final behavior is considered.

3. Conclusions

The harmonization of statistical techniques is an effective instrument in the analysis of econometric and psychometric systems. The study of behavior as a combination of psychosocial elements with economic and financial ones requires high mathematical rigor and consequently the utilization of the capacities of different sectors of statistical analysis is a useful finding. We have realized an integrated analysis of consumer reaction toward marketing tactics and strategies which could be generalized methodically for a larger area. We observe that "the state" of average expenses becomes more stationary after the discounts have been applied. Therefore, analyses of the market, measurements of quantities, and statistical study for this system should be better performed on the after-discounts states. Particularly we conclude that the consumer reaction to the discounts

was characterized by the increase in spending itself, not only the volumes of items purchased. We identified the load of each factor in the increase of expenditures and acknowledged the utility form in this case. The q-distribution analysis technique is quite successful because it constitutes a numerical measure or control test for the stability of the distribution and the stationary state itself. In the case of conditional purchase realism (Basic) we notice that the customer's behavior is rational, he weighs the alternatives in terms of his utility.
In this way, the buyer studied in this paper is (results) a dynamic agent. We have thus arrived at a more descriptive finding of the system, i.e. at a more natural form of its behavior as a result. In conclusion of this deductive analysis, we conclude that the results of the second stage in the attitude of the buyer in this case are more objective and can be generalized to other systems.

References

[1]. Gene V Glass Percy D. Peckham, James R. Sanders Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance . *Review of education research vol. 42, no. 3.* 1972
[2]. Elmira Kushta, Dode Prenga, Fatmir Memaj. Analysis of consumer behavior in a small size market unit: case study for Vlora District, Albania. IJSRM,2018
[3]. Mugo Fridah W.,Sampling in research.
[4]. Jorge Faber and Lilian Martins Fonseca. How sample size influences research outcomes. Dental Press J Orthod. 2014 Jul-Aug; 19(4): 27–29.
[5]. Ronald Jay Polland. Essentials of survey research and analysis. https://www.psycholosphere.com
[6]. Steiger, J.H. (1990), "Structural model evaluation and modification," Multivariate Behavioral Research, 25, 214-12.
[7]. Kalr Jorskog, Arthur Goldbwerg. Estimation of a model with multiple indicator and multiple causes of ingle latent variable. *Journal of the American statistical association. Volume 70, issue 351*(Sep. 1975).
[8]. Sabir Umarov, Constantino Tsallis, Murray Gell-Mann, Stanly Steinberg. Generalization of symmetric -stable Lévy distributions for q>1. *Journal of mathematical physics* 51, 033502 2010.
[9]. M.A. Robinson. Quantitative research principles and methods for human-focused research in engineering design. '*Research methods' publications*. May 2016. DOI: 10.1007/978-3-319-33781-4_3.
[10]. F. Shneider, R. Dell'Ano . Estimating the underground economy by using mimic models: a response to T. Breusch´s critique by dell'Ano , Working papër no. 0607. july 2006
[11]. Kwiatkowski, D.& al. (1992)..Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? . Journal of Econometrics 54 (1992) 159-178
[12]. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the çase of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(5), 157-175.
[13]. Steiger J.H (1990) ,"Structural model evaluation and modification," Multivariate Behavioral Research, 25, 214-12.
[14]. Khan, S., Memon, B. and Memon, M.A. (2019). Meta-analysis: a critical appraisal of the methodology, benefits and drawbacks. *Br J Hosp Med (Lond)*, 80(11), 636-641. doi:10.12968/hmed.2019.80.11.636
[15]. Constantino Tsallis Computational applications of non-extensive statistical mechanics. Journal of Computational and Applied Mathematics 227 (2009) pp 51-58.
[16]. Constantino Tsallis Computational applications of non-extensive statistical mechanics. Journal of Computational and Applied Mathematics 227 (2009) pp 51-58.
[17]. Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. Statistics in Medicine, 22, 1433–1446.
[18]. Steiger, J.H. (1990), "Structural model evaluation and modification," Multivariate Behavioral Research, 25, 214-12.
[19]. Bollen, K. (1989). Structural equation modeling with latent variables. New York, NY: Wiley.