Current Research Journal of Economic Theory 4(3): 77-82, 2012
ISSN: 2042-485X

Submitted: March 31, 2012 Accepted: April 17, 2012 Published: June 30, 2012

An Application of Ordered Logit Model and Artificial Neural Networks in an Income Model

[1]Ercan Baldemir, [2]Hatice Ozkoc, [3]Hakan Bakan and [4]Burak Yesildag
[1]Department of Business Administration, Mugla University, Mugla, Turkey
[2]Department of Statistics, Mugla University, Mugla, Turkey
[3]Department of Business Administration, Mugla University, Mugla, Turkey Mugla University, Mugla, Turkey

Abstract: The detection of factors affecting the income distribution is an economically and socially important study subject. Therefore, the researchers aim to model the income and by this way, make suggestions to overcome the difference. In this study, it is aimed to determine the factors affecting the income by using the results obtained from the study of “2010 A Study on Life Satisfaction”, which was performed by TUIK in Turkey. The income models, which were obtained with the help of artificial neural networks and ordered logit model, were compared in terms of estimation power and the obtained results were interpreted.

Keywords: Artificial neural network, estimation, income, life satisfaction, nodes, ordered logit model

INTRODUCTION

Personal income distribution expresses the distribution of income between individuals, families and consumer units, in accordance with age, occupation, gender and educational background. Personal income distribution is used as an important indicator of interpersonal income inequalities in economic, social and political areas to explain economic inequalities in different ways (Tunc, 1997). It is of great importance especially for policy makers to determine the factors that reveal the difference in income distribution and produce policies accordingly in an attempt to overcome the inequality. The studies of revealing the factors that determine the income differences is based on the model of human capital, which was initially presented by Mincer (1974). Being the primary element of the income difference, the model reveals the educational levels of individuals. In other studies that were performed on the income difference, the factors of human capital were revealed to be the primary reason of the difference, as well Ucdogruk *et al*. (2000). As a result of the studies performed on the relation between the income distribution and human capital, it was concluded that the income would be distributed in a more balanced way through education by focusing on education among the investments of human capital.

The objective of this study is to model the personal income model with artificial neural networks and ordered logit model by taking the theory of human capital into consideration and then compare the results. The artificial neural networks, which have a broad usage area today, model the system in accordance with the determined inputs and outputs without requiring any assumptions, in respect of its structure. The structures of input and output variables that are used in the phase of modelling with artificial neural networks determine the transfer function to be used in the model. Since the income variable that is used in the study is an ordered variable, as well as artificial neural networks, it could also be modelled with the help of ordered logit model, which is one of the discrete preference models. The data set was divided into two sections as education and test group for both of the methods and the accurate claffisication percentages were compared on the basis of models in both of the the sections.

METHODOLOGY

Ordered logit model: As discussed by Long (1997) the ordinal regression model was developed independently in the social sciences (in terms of an underlying latent variable with observed, ordered categories) and in biostatistics (where it is referred to as a proportional odds model).

In many economic applications, the dependent variable is discrete and represents an outcome of a choice between a finite set of alternatives. A number of qualitative response models deal with this characteristic of the dependent variable (Amemiya, 1981; Greene, 1996). Further, in some applications, there are multinomial choice variables that are naturally ordered. In this application, naturally ordered income variable is used as the dependent variable (Oksuzler, 2008).

Following to the relevant literature. Mincer (1974) earning equation is used in this study. The model that is to be estimated is as follows:

Corresponding Author: Hakan Bakan, Department of Business Administration, Mugla University, Mugla, Turkey

$$Income_i = \alpha_1 Education + \alpha_2 Age_i + \\ \alpha_3 Age_i^2 + \alpha_4 Sex_i + \varepsilon \quad i = 1, 2, ..., n$$

According to this model individual income is expected to be positively affected by individual's education level. Experience is measured by age and represented by al linear and a quadratic term to capture the nonlinearity in the earnings profile (Oksuzler, 2008). Age is used as a Proxy for experience. Sex variable is used to determine to gender effect on income. ε is the random error term. In order to compare the returns for different education levels education variable is splitted into dummy variables and the following model is development:

$$Income_i = \alpha_1\ HighSchool + \alpha_2 College + \alpha_3 University + \\ \alpha_4 Age_i + \alpha_5 Sex_i + \alpha_6\ MaritalStatus + \alpha_7\ Location + \varepsilon \quad i = 1,\ 2, ..., n$$

The ordered logit model is built around a latent regression where y_i* is the unobserved dependent variable, x a vector of explanatory variables, β an unknown parameter vector and ε the error term:

$$y_i^*\ \beta x_i + \varepsilon_i$$

Instead of y_i* the following is observed:

$$\begin{aligned} y &= 1\ \textit{if} && \mu_0 \leq y^* < \mu_1 \\ y &= 2\ \textit{if} && \mu_1 \leq y^* < \mu_2 \\ y &= 3\ \textit{if} && \mu_2 \leq y^* < \mu_3 \\ &\vdots && \\ y &= J\ \textit{if} && \mu_{j-1} \leq y^* \end{aligned}$$

where y is the category of income per annum ranked into 3 categories. u is the vector unknown threshold parameters estimated with the β vector. The ε is the disturbance term, which is assumed to be logistically distributed with mean 0, variance $\pi^2/3$ and cumulative distribution function:

$$F(\varepsilon) = \frac{\exp(\varepsilon)}{1 - \exp(\varepsilon)} = \frac{1}{1 - \exp(-\varepsilon)}$$

Consequently:

$$\Pr[y_i = j] = \Pr[y^* \text{ is in the } j \text{ th range}]$$

Hence the probability of observing an outcome may be written:

$$\Pr[y_i = j] = F[\mu_j - \beta x_i] - F[\mu_{j-1} - \beta x_i]$$

This implies that:

$$\Pr[y_i = j] = \frac{1}{1 + e^{-\mu_j + \beta x_i}} - \frac{1}{1 + e^{-\mu_{j-1} + \beta x_i}}$$

The above equation can be used to derive a likelihood function and subsequently maximum likelihood estimates of μ and β (Wooldridge, 2002).

Artificial neural networks: Classification is one of the most active research and application areas with regard to neural networks. The literature is vast and growing. Neural networks have emerged as an important tool for classification purposes. The recent research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods. The advantage of neural networks lies in the following theoretical aspects (Yay and Akıncı, 2009).

Artificial Neural Networks are systems that are deliberately constructed to make use of some organizational principle resembling those of human brain (Haydar *et al*., 2006). The Artificial Neural Network (ANN) model, as a data mining technique, has the capability to perform better than traditional statistical techniques. The strenght of ANN model lies in its ability to estimate non-linear and complicated processes without requiring a spesific assumption on either its input or output variables. This model allows computing, learning and remembering similar to human brains (Larasati *et al*., 2011).

Artificial Neural Networks (ANN) are relatively crude electronic networks of "neurons" based on the neural structure of the brain. They process records one at a time and "learn" by comparing their classification of the record (which, at the outset, is largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network and used to modify the networks algorithm the second time around and so on for many iterations (Yay and Akıncı, 2009).

The simplest building block of neural network consists of three layers: input, output and hidden layers. The first layer has one or more neurons (nodes) that represent independent variables, while the output layer consists of one more neurons that are dependent (outcome) variables. The output layer represents the model's classification deicisions, in which each decision class has one node. The hidden nodes in the model connect the input and output layers indirectly. In general, one or more hidden layers can be located between the input and output layers (Larasati *et al*., 2011).

An ANN has each processing element (the neuron) receiving inputs from the other elements, the inputs are weighted and added, the result is the transformed (by a transfer function) into the output (Fig. 1). The transfer function may be a step, sigmoid, or hyperbolic tangent function among others (Li, 1994). The data from the input neurons is propogated through the network via the interconnections such that every neuron in a layer is connected to every neuron in the adjacent layers. Each

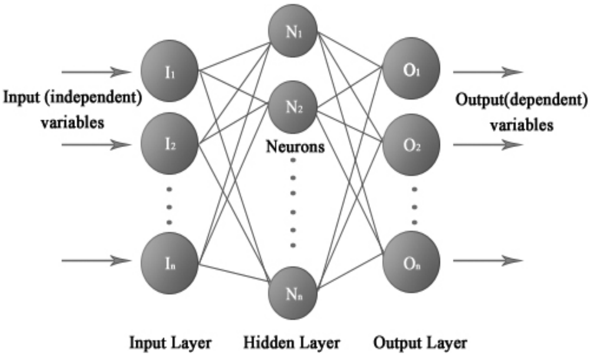

Fig. 1: The simplest building block of ANN

interconnection has associated with it a scalar weight, which acts to modify the strength of the signal passing through it. The neurons within the hidden layer perform two tasks: they sum the weighted inputs to the neuron and then pass the resulting summation through a nonlinear activation function (Haydar *et al*., 2006).

Learning or training is the term used to described the process of finding the values of these weights. The two types of learning associated with neural networks are supervised and unsupervised learning. Supervised learning occurs when there is a known target value associated with each input in the training set.the output of the network is compared with the target value and this difference is used to train the network. There are many differnet algorithms for training neural networks using supervised learning: backpropogation is one of the more common ones. Unsupervised learning is needed when the training data lack target output values corresponding to input patterns. This type of training is no source of feedback in the training process (Warner and Misra, 1996).

An ANN may have either a recurrent or nonrecurrent structure. A recurrent network is a feedback network in which the network calculates its outputs based on inputs and feeds them back to modify the inputs. As fort he noncurrent networks, data follow in one direction, from input layer to output layer without any feedback loop: they are also called feed forward networks. This type of networks has accounted for most existing ANN applications (Li, 1994). It is widely accepted that a three-layer feedforward network with an identity transfer function in the output unit and sigmoid functions in the middle-layer units can approximate any continuous function arbitrarily well given sufficient amount of middle-layer units (Alon *et al*., 2001).
The sigmoid activation function can be expressed as:

$$f(x) = \frac{1}{\left(1 + e^{(-n-x)}\right)}$$

n indicates the threshold (intercept point) and x indicates the aggregate of weighted value (Larasati *et al*., 2011).

The error correction learning procedure is simple enough in conception. The procedure is as follows: During training an input is put into the network and flows through the network generating a set of values on the output units. Then, the actual output is compared with the desired target and a match is computed. If the output and target match, no change is made to the net. However, if the output differs from the target a change must be made to some of the connections (Fuller, 1995).

Comparisons between ordinal logistic regression and artificial neural networks: Neural networks can automatically transform and represent highly complex nonlinear relationships more effectively than regression (Detienne *et al*., 2003). In ordinal logistic regression, the model complexity is already low, especially when no or few interaction terms and variable transformations are used. Performing variable selection is a way to reduce a model’s complexity and consequently decrease the risk of overfitting. Compared to logistic regression, neural network models are more flexible and thus more susceptible to overfitting (Dreiseitl and Ohno-Machado, 2002).

Neural network method can make rapid prediction once model are trained. The ability of learning on very large data set and predicting in time makes neural network method a useful and competitive tool for ordinal regression tasks (Cheng *et al*., 2008). Neural networks are data-driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form fort he underlying model (Yay and Akıncı, 2009).

The Ordinal Logistic Regression also has the capability to handle non-linear relationships among independent and dependent variables as the ANN model does, since it incorporates an exponential term in its function. The ANN model does not requires model specification known priori since the network has a capability to learn the relationship between layers based on the data pattern. Therefore, the ANN model provides more flexibility and higher robustness to model misspecification than the OLR model (Larasati *et al*., 2011).

The solution quality of an ANN is known to be affected by the number of neurons at each layer, the transfer function of each neuron and the size of the training set (Li, 1994). The OLR model requires the proportional odds assumption for ordinal data. This assumption implies that the cut point of spesific odd ratios are homogenous. The test of parallel lines is commonly used to determine whether thi assumption is fulfilled by ordinal data (Larasati *et al*., 2011).

The contribution of parameters in ordinal logistic regression can be interpreted, whereas this is not always the case with the parameters of a neural network (weights) (Dreiseitl and Ohno-Machado, 2002). For ordinal logistic regression, the popularity may be attributed to the interpretability of model parameters and ease of use; for artificial neural networks, this may be due to the fact that these models can be seen as nonlinear generalizations of logistic regression and thus at least as powerful as that model (Dreiseitl and Ohno-Machado, 2002).

The neural network approach to dealing with missing values in a data set and to reconstructing a data set is more efficient than conventional regression-based methods. Compared to ordinal logistic regression, neural networks deal with both linear and nonlinear data, formulate the correct data model without a priori specification by the researcher, require less stringent assumptions than regression, learn from experince, see through noise and irrelevant data, offer a high degree of robustness, perform well with limited data, deal affectively with missing data, combine two or more models at the same time and complete real time applications (Detienne *et al*., 2003).

CASE STUDY

The data set of the "Study on Life Satisfaction", which was performed by TUIK (2010), was used for the purpose of comparing the artificial neural networks and ordered logit model, in terms of explaining the income model. Since the income variable was ordered, the econometric model to be used was determined as the ordered logit model. In the study, gender, marital status, location and educational level were included in the model as the variables, which were considered to explain the income-dependent variable.

The definitive statistics of the dependent and independent variables are given in Table 1. As is seen in Table 1, taking the income level into consideration, it is seen that a great majority of answerers (45.92%) has a low income level. On the other hand, the rate of those who have the highest income level is very low. The gender distribution is almost equal on the data set and the rate of unmarried individuals was determined to be higher than the married individuals. Examining the distribution of survey participants in terms of their educational levels, the rate of those who did not receive education after the primary education was found to be very high (71.75%).

Table 1: Descriptive statistics

Variables	Frequency	Percent
Dependent variable: Income		
0-990 TL	3186	45.92
991-1650 TL	2074	29.89
1651 and above	1678	24.19
Gender		
Male	2989	42.95
Female	3958	57.05
Marital status		
Married	1677	24.17
Not married	5261	75.83
Location		
Urban	4935	71.13
Rural	2003	28.87
Education		
Primary education	4978	71.75
High school	1194	17.21
College	261	3.76
University	505	7.28

Table 2: Actual ıncome level*predicted ıncome level for ANN (train)

	Predicted level of ıncome			
Actual level of ıncome	0-990 TL	991-1650 TL	1651 TL-	total
0-990 TL	1007	1404	43	2454
	41%	57.2%	1.7%	100%
991-1650 TL	299	1318	70	1687
	17.7%	78.1%	4.1%	100%
1651 TL and above	93	1075	241	1409
	6.6%	76.2%	17.1%	100%
Total	1399	3797	354	5550
	25.2%	68.4%	6.3%	100%

Table 3: Actual ıncome level*predicted ıncome level for ANN (test)

	Predicted level of ıncome			
Actual level of ıncome	0-990 TL	991-1650 TL	1651 TL-	total
0-990 TL	405	318	9	732
	55.3%	43.4%	1.2%	100%
991-1650 TL	141	230	16	387
	36.4%	59.4%	4.1%	100%
1651 TL and above	45	168	56	269
	16.7%	62.4%	20.8%	100%
Total	591	716	81	1388
	42.5%	51.5%	5.8%	100%

During the rest of the study, the income model will initially be modelled with artificial neural networks and then the same model will be constituted with the ordered logit model and both of the models will be compared in terms of estimation powers.

Examining the data in artificial neural networks, best predictions are taken with three layered structure using Levenberg Marquardt training algorithm. First layer is the input layer which contains information about gender, age, location, marital status and education situation as our input variables. These variables are computed in the hidden layer for generating predictions at the output layer classifying between three income intervals. After some experimentation it was found that using ten neurons in the hidden layer gives the most accurate results for this data set (Fig. 2).

The data set contains 6948 values with seven input variables and corresponding income intervals. Eighty percent of data which consists of 5550 experiments is used for training the network and the 20% that consists of 1398 experiments is used for testing the trained network comparing the values of the actual and the predicted

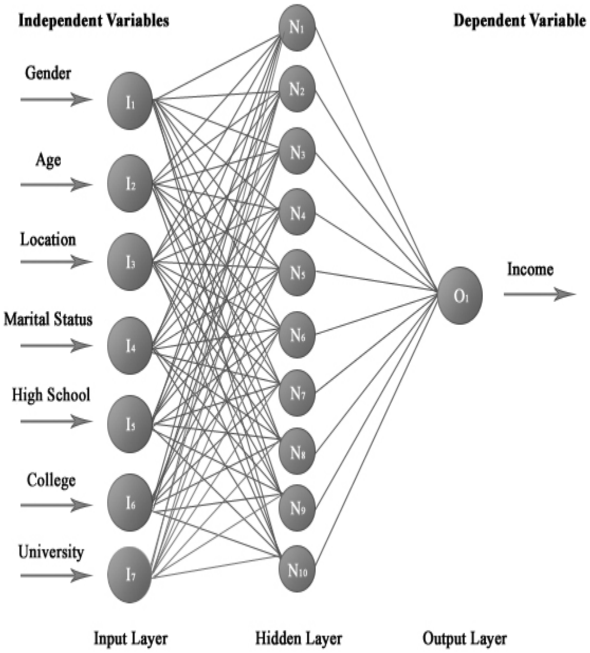

Fig. 2: Artificial neural network used in the study

income interval. The network is best trained at 320 iterations with these circumstances. After training, network is simulated to get predictions on income intervals for the seperated test data and 49.78% accuracy is reached. The same network is also simulated to predict income intervals for the training data providing 46.23% accuracy. The detailed values taken by network simulation are given Table 2 and Table 3 containing actual response intervals and corresponding predicted intervals for each one.

During the rest of the study, the income-dependent variable was modelled with the help of the ordered logit model. Examining the results given in Table 4, it was observed that the marital status and gender do not have a statistically significant effect upon the income. According to one of the results obtained from the model, the income increases exponentially together with the increase of the educational level. Besides, according to the results of the ordered logit model, the income levels of city dwellers were found to be higher than the levels of those living in rural areas.

Five thousand five hundred and fifty observations of the data set were modelled with the help of OLM, in an attempt to determine the estimation accuracy of the ordered logit model on the dependent variable for education. As a result of the analysis, the accurate values and estimated values of the dependent variable are given in Table 5. It is seen that 34.9% of 5550 observations were assigned to the accurate income group at the end of the modelling. This rate is considerably lower than the rate obtained with ANN. Additionally, the model results were adapted into the remaining 1388 observations in the data set and by this way, the estimation values were obtained for the income variable.

Table 4: Ordered logit model estimation results

Dependent variable: Income (1: 0-990 Tl, 2: 991-1650, 3:1651 and above)	$\hat{b}$	Std. error	Z	p>|z|	exp($\hat{b}$)
Age (A)	-0.0001	0.0016	-0.07	0.947	0.999
Gender					
Male (M)	-0.0638	0.0489	-1.30	0.193	0.938
Marital status					
Married (MRD)	-0.3611	0.0579	-6.23	0.000*	0.697
Education					
High school	1.1648	0.0655	17.79	0.000*	3.205
College	1.8972	0.1262	15.03	0.000*	6.667
Universtiy	2.9125	0.1149	25.34	0.000*	18.403
Location					
Urban	1.0108	0.0565	17.87	0.000*	2.747
/cut1	0.8244	0.0919			
/cut2	2.4388	0.0965			
Log likelihood = -6476.4062LR chi2(7) = 1778.58					
Prob>chi2 = 0.0000*Pseudo R^2 = 0.1207					

*: Coefficient is statistically significant at a significance level of 5%; Base categories: Female, Not married, Primary education, Rural

Table 5: Actual ıncome level*predicted ıncome level for OLM (train)

	Predicted response ınterval			
Actual level of ıncome	0-990 TL	991-1650 TL	1651 TL-	total
0-990 TL470.0	128.0	44.0	642.0	
	73.2%	19.9%	6.9%	11.6%
991-1650TL	592.0	208.0	103.0	903.0
	65.6%	23.0%	11.4%	16.3%
1651 TL and above	1392.0	1351.0	1262.0	4005.0
	34.8%	33.7%	31.5%	72.2%
Total	2454.0	1687.0	1409.0	5550.0
	44.2%	30.4%	25.4%	100%

Table 6: Actual ıncome level*predicted ıncome level for OLM (test)

	Predicted response ınterval			
Actual level of ıncome	0-990 TL	990-1650 TL	1650 TL-	total
0-990 TL	231.0	46.0	23.0	300.0
	77.0%	15.3%	7.7%	21.6%
990-1650 TL	217.0	127.0	49.0	393.0
	55.2%	32.3%	12.5	28.3%
1650 TL-	284.0	214.0	197.0	695.0
	40.9%	30.8%	28.3%	50.1%
Total	732.0	387.0	269.0	1388.0
	52.7%	27.9%	19.4%	100%

Table 6 includes the results obtained for the test group. The general accuracy percentage of the test group was determined to be 39.9%. As well as the education group, the test group obtained a higher success percentage in modelling with NN than the modelling with OLM.

RESULTS

In this study, the income model was modelled with the help of artificial neural networks and ordered logit model and the accuracy percentages of both models on estimation were compared.

According to the results obtained, artificial neural networks attained a greater accuracy percentage compared to the ordered logit model in both education and test groups. In that case, it was concluded that artificial neural networks explain the income model much better.

However, as it is already explained in previous sections, both of the models certainly have advantages and disadvantages. While the explanatory variables affecting the income variable that is modelled with the help of the ordered logit model are revealed as a result of the model, no information is obtained in artificial neural networks regarding the significance of independent variables. Additionally, the interpretation of the mathematical model that is obtained with artificial neural networks is harder and more complicated than the ordered logit model. Apart from these advantages, the ordered logit model displayed a lower performance in the estimation of the dependent variable, compared to the artificial neural networks. Besides, since the artificial neural networks do not include assumptions required by the econometric models, it is a more useful approach than the ordered logit model.

REFERENCES

Alon, I., M. Qi and R. Sadowski, 2001. Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. J. Retail. Consumer Serv., 8: 147-156.

Amemiya, K.,1981. Qualitative response models: A survey. J. Econ. Literat., 194: 483-536.

Cheng, J., Z. Wang and G. Pollastri, 2008. A neural network approach to ordinal regression. 2008 International Joint Conference on Neural Networks, pp: 1279-1284.

Detienne, K., D.H. Detienne and S.A. Joshi, 2003. Neural networks as statistical tools for business researchers. Organ. Res. Method., 6(2): 236-265.

Dreiseitl, S. and L. Ohno-Machado, 2002. Logistic regression and artificial neural network classification models: A methodology review. J. Biomed. Inform., 35(5-6): 352-359.

Fuller, R., 1995. Neural Fuzzy Systems. Abo Akademi University, Abo, Retrieved from: http: //faculty. petra.ac.id/ resmana/ basiclab/ fuzzy/fuzzy_book. pdf, (Accessed on: January 2012).

Greene, W.H., 1996. Econometric Analysis. Prentice Hall, Englewood Cliffs, N.J.

Haydar, A., Z. Ağdelen and P. Ozbeşeker, 2006. The use of back-propagation algorithm in the estimation of firm performance. Istanbul Ticaret Universitesi Fen Bilimleri Dergisi, 5(10): 51-64.

Larasati, A., C. De Yong and L. Slevitch, 2011. Comparing neural network and ordinal logistic regression to analyze attitude responses. Serv. Sci., 3(4): 304-312.

Li, E.Y., 1994. Artificial networks and their business applications. Inform. Manage., 27(5): 303-313.

Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variables. Sage Publication, Thousand Oaks.

Mincer, J., 1974. Schooling, Experience and Earning. National Bureau of Economic Research, United States of America.

Oksuzler, O., 2008. Munich Personal RePEc. Paper No. 14375, Retrieved from: http://mpra.ub.uni-muenchen. de/14375/pdf. (Accessed on: January 2012).

Tunc, M., 1997. Size Analysis of Human Capital Development Human Capital Approach and Turkey. DEU. Sosyal Bilimler Enstitüsü, Izmir, (In Turkish).

TUIK, 2010. Study on Life Satisfaction, Ankara, Turkey.

Warner, B. and M. Misra, 1996. Understanding neural networks as statistical tools. Am. Statist., 50(4): 284-293.

Wooldridge, J.M., 2002. Econometric Analysis. MIT Press, Cambridge, Mass.

Ucdogruk, S., M. Ozcan and Z. Ozcan, 2000. Indices of the factors determine differences in selected provinces by mean of development in Turkey. Ekonomik Yaklasım, 11(23): 29-57, (In Turkish).

Yay, M. and E.D. Akıncı, 2009. Application of ordinal logistic regression and artificial neural networks in a study of student satisfaction. Cypriot J. Educ. Sci., 4: 58-69.