Application of Ordered Logit Model in Investigating the factors Affecting People's Income (A Case Study in Tehran City)

Hosein Mohammadi[1], Sasan Torabi[2], Azadeh Dogani[2]

[1]Academic Member of Agricultural Economics Department, Ferdowsi University of Mashhad, Mashhad, Iran, mohammadi111@gmail.com
[2]Ph.D Student of Agricultural Economics, International Campus of Ferdowsi University of Mashhad, Mashhad, Iran, sassan_torabi@yahoo.com
Corresponding Author: Hosein Mohammadi

Abstract
The study aimed at investigating the factors affecting people's income in Tehran city. The data was collected through field survey, designing and running a questionnaire using simple sampling method in 2013. The factors affecting income were estimated using ordered logit based on maximum likelihood (ML).Income was evaluated as a function of education, age, age squared, gender, residence, occupation, wealth and private investment variables. The results indicated that the increase in the age (continuous variable), gender (for women), residence (for non-capital of the provinces), wealth (for those who had not inherited) and private investment (for those who had not invested) decreases the likelihood of the people to be placed at higher income levels, and increase in the age squared variable increases the likelihood of the people to be placed at higher levels of income. In other words, people's income increases from certain age onwards. In case of education and occupation, Increase in the levels of these variables increases the likelihood of people to be placed at higher income levels. The calculated marginal effects are also in accordance with the theoretical expectations. Based on calculated Pseudo R-Square statistics, it can be said that estimated ordered logit model also enjoys a high level of goodness of estimate.
Keywords: Ordered Logit Model, Income, Tehran, ML Estimation Method

Introduction
Income disparities issue has long been existed among the different communities and different groups of people of societies. Searching for the causes of this problem and for ways to solve it has been a concern for economists and political leaders. In recent decades, the emergence of a new branch of economics called the Economics of Education has provided a new explanation for the differences in income in different societies. The economists have been studying and researching about it using the concept of human capital. Based on this concept,

Future economic benefits of education, particularly its income benefits, is among the most important factors for people in decision making to take part in educational activities.

Concurrent with the formation and evolution of science of economics, attention to education in the economics literature, has attracted the attention of a wide range of economists. Many of classical, neoclassical and contemporary economists have presented ideas and conducted empirical studies in this area. Most of the empirical studies in this area have been done (by people like Schultz and Becker) after the development of theory of human capital in economics and Earning Function is also developed by Mincer (Psacharopoulos, 2002). The Mincer earning function model for estimating the rate of return to education, is one the models that has been widely used by researchers and scientists. The model estimates the effect of education on income based on the classical econometric rates.

In addition to education, there are other factors that affect people's income such as age, gender, etc., which have been estimated as a function of education ,age, age squared, gender, residence, wealth and private investment variables. The innovation of the present study includes using nonlinear regression model with dependent discrete, categorical and limited variables. The paper has been developed in five sections. Second Section of the paper is dedicated to Mincer Rarning Function , third section to materials and methods, fourth section to presenting variables and the estimation of the model and fifth section to conclusion. The paper ends with presenting references and computational appendices.

Literature Review
The general results of some the studies that have explicitly or implicitly been conducted in connection with the topic of the present paper is presented here, so they can be used as guidelines in other sections of the study. Below are some of the conducted empirical studies. Hashemiyan (1991) in his MA thesis titled as 'The income effect of education' in Isfahan University, directly estimated the income functions of Iranian people. He estimated the income function using the sampling data obtained from textile, electricity and water facilities workers in Isfahan. The results of the study indicated a positive and significant effect of schooling and education on people's income, so it can be said with certainty that there is a positive and direct correlation between people's education level and the growth in their future income. Tansel (1994), Dayioglu and Kasnakoglu (1997) used the data from the Turkish Statistical Institute and applied Probit model. Both of the studies showed that education had a positive effect on income. Tansel indicated that elementary and high school education are more effective in women compared to men. This result is confirmed by Dayioglu and Kasnakoglu too. They found that education increased women's participation in the labor force. The studies indicated th positive marginal effect of education on income.

Chan and Kent (1999) investigated the estimation of Mincer earning function in Hong Kong. This study showed an increase in the rate of return and decrease in return of working experience for the years studied. They stated that due to structural changes and economic reforms in Hong Kong during recent years, people gradually came to the conclusion that just having the experience is not enough to ensure a good job and the role of education in people's income level is more visible. The model used in this study is as follows.

$$Ln(y_t) = \beta_\circ + \beta_1 S + \beta_2 S^2 + \beta_3 t + \beta_4 t^2 \qquad (1)$$

's' and 't' variables are, respectively, the years of education and years of experience. S^2 is included in the model to show nonlinear effects of education on income. The results of their study suggests that β_0 was declining during the years studied which indicates that Hong Kong market paid less to illiterates and those with low education over time. The calculated value

of β_2 is positive and statistically significant which is a strong proof for increasing rate of return to education and schooling; β_4 coefficient is negative and decreasing which proves that on the job training has a decreasing return which shows the depreciative effect of work experience on people's income.

Naderi (1999) in his doctoral dissertation at the University of London used multilevel models to study the impact of education on people's income in Iran. Using extended Mincer earning function and applying it in multi levels on data of manufacturing industry in Iran, he argued that education and work experience significantly and systematically correlate with workers' income.

Wolter and Weber's study (2000) in Switzerland shows that return to education of women is slightly more than that of men. But return to work experience is less for them. In addition, age-income curve for women is flatter than that for men.

Sari(2002), estimated the Mincer earning function using a dataset obtained from blue city in Turky. He showed that return to education for one year was 12.1 % and return to experience was 9.3%. He also showed that the return to a year of education had the highest value in elementary school and the lowest in high school.

Elmi (2002) in his doctoral dissertation titled as " The effect of human capital and governmental investment on human resources in the context of endogenous growth models in Iran" at Tehran university, studied the variables of annual income, squared years of experience, gender and literacy level of employees using extended Mincer earning function . The results of 34 sectional estimations between 1984 and 2000 for urban and rural areas of Iran indicated that people's income was affected by their level of literacy, age and years of experience.

Ozcan et al.(2013) investigated the relation between income from education in paid workers and business owners in Istanbul. They found that return to education for those who run their own business is higher than wage earners.

Materials and Methods
Data
In this study, simple stratified sampling method was used in order to select samples. In stratified sampling, the units of the studied population are grouped into categories which are more homogeneous in variable characteristics, so the changes are reduced within categories. Then, some samples are randomly selected from each category. Tehran city has 22 districts and in this study the categories were selected according on these districts. Data used in this research were collected through field surveys in 2013. To determine the number of samples, a pre-study was conducted. Firstly, 50 individulas were selected for the pre-study. The results of the pre-study showed that the studied characteristic variance was 0.616, so by applying relation 2 which determines the number of samples using simple stratified method, the number of samples was determined.

Relation 2 : $n = Z^2 (\varepsilon W_i \sigma_i^2) / d^2$ (2)

Where 'Z' is a normal value corresponding to a confidence level of α-1, 'd' the permissible error,' δ_i^2 ' is the studied characteristic and 'W$_{i'}$ is the weight of unit 'i'. Considering the abovementioned relation and the data obtained from the pre-study and $d_1^2 = 0.04$ and $\delta_1^2 = 0.616$, the number of samples was determined as 282 people. After extracting data and using ordered logit regression model, the effect of education on income was investigated. All statistical analyses and tests were done using SPSS package and Sata softwares.

Ordered Logit Model

Ordered logit model is based on a continues latent variable. The study used ordered logit model to investigate and determine the effect of descriptive variables of education, age, age squared, gender, residence occupation, wealth and private investment on people's income and impact of each variable on the likelihood of individuals to be placed at one of four income groups of 0 to 638, 638 to 1500, 1500 to 3000 and more than 3000 thousand Toman.
The model is as follows:

$$y_i^* = \beta' x_i + \varepsilon_i \qquad -\infty < y_i^* < -\infty \tag{3}$$

Where
y_i^* : income
β_i: Vector of parameters that should be estimated
x_i: Observed vector of non-random explanatory variable which shows the characteristic of ith person
ε_i: Residual error which is logistically distributed.
Since y_i^* is a latent variable, standard regression techniques are not applicable to estimate the sample size.
If y_i is considered as a discrete and observable variable which shows different levels of people's income, the relation between latent variable y_i^* and observable variable y_i is obtained from
ordered logit model as follows:

$$
\begin{aligned}
y_i &= 1 && \text{if} && -\infty \le y_i^* < \mu_1, && i = 1, \ldots, n, \\
y_i &= 2 && \text{if} && \mu_1 \le y_i^* < \mu_2, && i = 1, \ldots, n, \\
y_i &= 3 && \text{if} && \mu_2 \le y_i^* < \mu_3, && i = 1, \ldots, n, \\
\ldots && \ldots && \ldots && && \ldots \\
y_i &= J && \text{if} && \mu_{J-1} \le y_i^* < +\infty, && i = 1, \ldots, n,
\end{aligned}
\tag{4}
$$

In which 'n' is the value for the sample size.
'μ' and's' are the thresholds that define observed discrete answers and should be estimated.
The probability of yi =j should be calculated by the following relation:

$$\Pr(y_1 = J) = \Pr(y_1 \ge \mu_{J-1}) = \Pr(\varepsilon_1 \ge \mu_{n-1} - \beta x_1) \tag{5}$$
$$= F(\beta x_1 - \mu_{J-1})$$

In cumulative probability expression, ordered logit model estimates the likelihood of person 'I' to be at 'jth' level or less (1,…,j-1). It should be noted that the answer groups in ordered logit model are ordered.
Ordered logit model is expressed as follows:

$$\log\left[\frac{\gamma_j(x_i)}{1-\gamma_j(x_i)}\right] = \mu_j - [\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}] \tag{6}$$

$$j = 1, \ldots, J; i = 1, \ldots, n$$

In which 'Υ $_j$' cumulative probability is as following:

$$\gamma_j(x_i) = \gamma\left(\mu_j - \beta' x_i\right) = P\left(y_i \le j \mid x_i\right) \tag{7}$$

B$_i$ is the column vector and of (β$_1$, β$_1$,……,β$_k$) parameters and xi is the column vector of explanatory variables. . μ$_j$ is only dependent on probability of predicting category and is not dependent on explanatory variables

Furthermore, the crisp part

$$\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} \tag{8}$$

is the independent part of the category. These two characteristics ensure that the answers groups are ordered and show that the results are a series of parallel lines.

In order to evaluate the hypothesis of equality of the parameters for all the groups, parallel regression test is used. This test compares the estimated model with a series of coefficients for all the groups with a model with a separate series of coefficients for each group. In this case, if the current model, which is the null hypothesis, is accepted, it proves that the status parameters are the same for all the answer groups.

X^2 statistic in parallel regression test is calculated as follows:

$$\chi^2 = -2\text{LogLikelihood}_{Cm} - (-2\text{LogLikelihood}_{Gm}) \tag{9}$$

In which, G m and C m represent the current and general model respectively. If the calculated X^2 in more than X^2 in the table, it will indicate the rejection of the null hypothesis which means the current model is estimated correctly. Parameters are estimated by maximum likelihood estimation method, which maximizes the probability of categorization.

$$L\left(y \mid \beta; \mu_1, \mu_2, ..., \mu_{J-1}\right) = \prod_{i=1}^{n} \prod_{j=0}^{J} \left[\gamma\left(\mu_j - \beta' x_i\right) - \gamma\left(\mu_{j-1} - \beta' x_i\right)\right]^{Z_{ij}} \tag{10}$$

Where 'z_{ij}' is a binary variable. It equals to 1 when the observed group for person 'i' is 'j' ,and if they are not equal, it equals to zero. Newton-Raphson algorithm is used in the maximizing process.

In this model, the interpretation of R^2-pseudo is different and it does not have the typical interpretation. With the increase of the estimates of the model, its value also increases and Its value is between zero and one. Berman and Benaque in 1985 introduced the classification accuracy to evaluate the goodness of estimates of the model, which expresses the percentage of correct and incorrect predictions of the dependent variable.

Regarding the interpretation of the coefficients, it should be noted that coefficients are not directly interpreted in this model. In case the predicting variable increases, changes in probability are dependent on two factors: one of them is the predicting value and the other is dependent on other variables; considering the fact that changes in probability are not constant, coefficients are not directly interpreted.Thus, marks are used to change the probability in this model for the extreme groups (first and last). For instance, the positive mark next to β_k coefficient indicates that ∂x_k increase in predicting variable ,decreases the probability of the first category (yi=1),while the probability of the last group (yi=j) increases, each act in the opposite direction direction.in such circumstances, the directions of the middle categories are not clear.

The calculation of the marginal effect of one unit change in X_k predictor on the probability of 'j' category is as follows:

$$\frac{\partial P\left(y_i = j \mid x_i\right)}{\partial x_k} = \left[\frac{\partial \gamma\left(\mu_j - \beta' x_i\right)}{\partial x_k} - \frac{\partial \gamma\left(\mu_{j-1} - \beta' x_i\right)}{\partial x_k}\right] \tag{11}$$

$$= \left[\lambda\left(\mu_{j-1} - \beta' x_i\right) - \lambda\left(\mu_j - \beta' x_i\right)\right]\beta_k$$

In which $\mu_j = +\infty$, $\mu_o = -\infty$, $\lambda_j(x_i) = \frac{\partial \gamma_j(x_i)}{\partial x_k}$.(Maddala)

Making decisions about using variables' value in estimation is very important, because the marginal effect depends on the values of all explanatory variables. Since total Probability always equals 1, the total marginal effect for each variable is zero. But it should be noted that the marginal effect is not direct on binary variables, and it can obtained by calculating the difference between the two possible probabilities.

Results and Discussion
The Introduction of Variables, Model Estimation And Data
Table 1 briefly specifies the variables used in the estimated model

Table 1
Description of the Variables Used in the Estimated Model

Variable	Description
1.Income(Dependent variable)	Zero to 638 (Thousand Tomans) = 1 638 to 1500 (Thousand Tomans) = 2 1500 to 3000 (Thousand Tomans)= 3 more than 3000 (Thousand Tomans) = 4
2.Age	continuous
3. $(Age)^2$	continuous
4.Gender	Women=0 Men=1
5.Education	High school diploma and lower=1 Associate and undergraduate degree=2 Graduate degree= 3 Ph.D = 4
6.Residence	Capital of the province = 1 Non-capital of the province=0
7.Occupation	Government job = 0 Non-government job=1 Jobless=2
8.Wealth	Have inherited=1 Have not inherited=0
9.investment	Yes=1 No=0

Obtained from the research findings

To evaluate the effect of education on income and considering the ordered nature of the income groups, ordered Logit model was applied with the dependent variable of different levels of income, and the results are presented in Table 2.

Table 2			
Results of The ordered Model Estimation			
Variable	Coefficient	Standard error	Z statistic
Age squared	0.034	0.636	0.05

Age		-0.008	0.013	-0.62
Gender: woman-man(base)		-0.174	0.229	-0.76
Education	High school diploma and lower	-0.585	0.722	-0.81
	Associate and Undergraduate degree	-0.820	0.704	-1.16
	Graduate degree	0.424	0.677	0.63
	Ph.D(base)	-		
Residence	Non-capital of the province	-0.116	0.370	0.32
	Capital of the province(base)	-		
Occupation	Governmental	0.331	0.503	0.66
	Non-governmental	1.329	0.631	2.11
	Jobless(base)	-		
Wealth	Have not inherited	-3.032	0.400	-7.56
	Have inherited(base)	-		
Investment	No	-0.771	0.242	-3.19
	Yes(base)	-		

The source of research findings

According to calculated Pseudo R-Square statistics in Table3, it can be said that the estimated ordered logit model has a high level of goodness of estimates and independent variables used in the model describe a large proportion of changes in income levels.

Table 3
Pseudo R-Square

Cox and Snell	0.335
Nagelkerke	0.358
McFadden	0.149

Obtained from the research findings

Classification accuracy which shows the percentage of correct prediction of the model and which is presented as an alternative to Pseudo R- Square shows the high accuracy of the categorization of the model. Accordingly, the estimated ordered logit model correctly predicted about 60 percent of the changes in different levels of the income.

Due to some reasons such as using incorrect correlation function, using wrong model and wrong order of different dependent categories, the general model may have a significant improvement in estimation compared to the current model. Parallel regression test is used to investigate this matter.

The results of the test , which are presented in Table 4 ,suggest the rationality of the hypothesis of equality of parameters in all the categories of the estimated model. Considering

the significance level of x^2 statistic in parallel regression test, we can assume that the value of the status parameters for all answer groups are the same and fixed and so the estimation of the ordered logit model enjoys strong bases.

Table 4 the results of parallel regression test			
Model	-2Log Likelihood	Chi-Square	Significance level
Current model	494.054	71.341	0.720
General model	433.713		

Obtained from the research findings

Table 5, goodness of estimates, includes Pearson and Deviance tests with the null hypothesis of good estimation of the data by the model, so X^2 statistic calculated by Pearson and Deviance tests should confirm the validity of the null hypothesis.

Table 5.
Indicators of goodness of estimates

Statistic	Chi-Square	Significance level
Pearson	480.509	0.347
Deviance	394.601	0.995

Obtained from the research findings

Considering the significance level of calculated X^2 by these tests, the data are estimated properly by the model. Considering the tests and abovementioned information, the estimated model is sufficiently reliable and the results of the model are greatly reliable.

As can be seen in table 2, the variables of age (continuous), gender(for women), residence(for non-capital of provinces),wealth(fro those who have not inherited anything) and investment (for those who have not invested anything) affect income in negative direction(opposite direction). In other words, the increase in independent variables reduces the likelihood of the people be placed at higher levels of income. Also, according to the results presented in table 2, the variable of age squared has a positive effect on the level of income and the increase in this independent variable, increases the likelihood of the people be placed at higher levels of income. In other words, people's income increases from a certain age onwards.

In the case of education variable, the negative mark of coefficients indicate that the lower levels of education (high school diploma and lower, associate and graduate degree) have negatively affect the level of income (opposite direction direction) and higher levels of education (graduate degree and Ph.D) positively affect the level of income. In other words, the increase in education level, increases the likelihood of people to be placed at higher levels of income. Occupation variable positively affects income levels and increase in the level of this independent variable increases the likelihood of people to be placed at higher levels of income.

The marks for estimated coefficients only can be interpreted regarding which income category they can be placed. Therefore, in order to derive further results from the estimated model, marginal effects should be calculated or each income category. For this purpose, the marginal effect for each income category is calculated and the results are presented in Table 6.

Table 6
Marginal effects for different income levels

Variable		Level1	Level2	Level3	Level4
Age squared		-0.004	-0.001	0.002	0.003
Age		0.001	0.0003	-0.0006	-0.0009
Gender: woman-man(base)		0.026	0.006	-0.0138	-0.018
Education	High school diploma and lower	0.077	0.037	-0.045	-0.069
	Associate and Undergraduate degree	0.115	0.044	-0.067	-0.092
	Graduate degree	-0.043	-0.037	0.018	0.61
	Ph.D(base)	-			
Residence	Non-capital of the province	0.017	0.004	-0.009	-0.012
	Capital of the province(base)	-			
Occupation	Governmental	-0.052	-0.008	0.028	0.032
	Non-governmental	-0.172	-0.078	0.088	0.162
	Jobless(base)			-	
Wealth	Have not inherited	0.351	0.186	-0.105	-0.433
	Have inherited(base)	-			
Investment	No	0.114	0.036	-0.066	-0.084
	Yes(base)	-			

The source of research findings

According to the results presented in table 6, with the increase in variables of age (continuous), gender(for women), residence(for non-capital of the province),wealth(for those who have not inherited) and private investment(for those who have not invested), the likelihood of people to be placed at lower levels of income increases and being placed at higher levels of income decreases.

Also according to the results, with the increase in average of age squared, the likelihood of people being placed at lower levels of income decreases and on the other hand the likelihood of them being placed at higher levels of income increases.

Moreover, according to the results in Table 6 on the education variable, the likelihood of people with lower levels of education (high school diploma and lower, associate and undergraduate degree) to be place at lower income levels increases and on the other hand the likelihood of them to be placed at higher income levels decreases and the likelihood of people with higher levels of education (graduate degree,PhD) to be place at lower income

levels decreases and on the other hand the likelihood of them to be placed at higher income levels increases.

Regarding occupation variable, the increase of jobs decreases the likelihood of the people to be placed at lower income levels and on the other hand the likelihood of them to be place at higher income levels increases. Increase in governmental employment increases income, but the increase in non-governmental occupations increases people's income more.

Conclusion

Considering the calculations and abovementioned descriptions,the following results are achieved.

1. The variables age (continuous), gender (for women), residence (for non-capital of the province),wealth(for those who have not inherited) and private investment(for those who have not invested) affect income level in negative direction(opposite direction). In other words, the increase in these independent variables decreases the likelihood of people to be placed at higher income levels. Considering the results of marginal effects, it can be said that with the increase in these independent variables, the likelihood of these individual to be placed at lower income levels increases and on the other hand the likelihood of them to be placed at high income levels decreases.

2. Age squared and occupation variable affect income levels in positive direction the increase in levels of these variables increases the likelihood of people to be places at higher levels of income levels. Also according to the results of marginal effects, the increase in level of independent variables, the likelihood of people to be placed at lower levels of income decreases and on the other hand the likelihood of people to be placed at higher levels of income increases.

3. Regarding the education variable, coefficients show that the lower levels of education (high school diploma and lower,undergraduate degree) affects income levels in negative direction (opposite direction direction) and high levels of education (graduate degree and Ph.D) affect income levels in positive direction.in other words, the increase in education level increases the likelihood of people to be placed at higher income levels. According to the results of the marginal effects on education variable,the likelihood of people with low levels of education(high school diploma and lower, undergraduate degree) to be placed at low income levels increases and on the other hand, the likelihood of the people to be placed at high income levels decreases. It also can be observed that the likelihood of people with graduate and post graduate degree to be at low income levels decreases and on the other hand the likelihood of these people to be placed at high income levels increases.

4. Based on the calculated Pseudo R-Square statistic, it can be said that the estimated ordered logit model enjoys a high level of goodness of estimates and the independent variables used in the model describe a high level of changes of income. Also, accuracy of categorization which shows the percentage of correct prediction of the model and is introduced as an alternative to Pseudo R-Square, shows the high accuracy of the applied model. Accordingly, the estimated ordered logit model correctly predicted about 60 percent of the changes in different levels of the income.

References

Chan, T., S. Kent., S. Lam. (2000) Estimation of the Human Capital Earning Funtion in Hong Kong. Hong Kong University of Science&Technology.Econ 347 Hong Kong Economy. Term Paper.PP:1-14

Elmi, Z. (2002). The effect of human capital and governmental investment on human resources in the context of endogenous growth models.Doctoral dissertation.Rghtesad faculty. Tehran university.

Green, W. (1997) Econometric Analysis. Upper Saddle River: Prentice Hall.

Hashemiyan Isfahani, Masood. (1991) The income effect of education. Master Thesis. Faculty of Administrative Sciences and Economics, Isfahan University.

Katz, L. and Autor, D., (1999) Changes in the Wage Structure and Earning Inequality, Orley Afshefelter and David Card (drl), Handbook of Labour Economics, Volume 3A, Amsterdam and New York:North-Holland

Krueger, A.B. and Lindahl, M., (1999) Education for Growth in Sweden and The World, NBER Working Papers No.7190, National Bureau of Economic Research.

Naderi, A. (2002). Multilevel models and dissonance and efficiency of budget in selected universities. Quarterly Journal of Research and Planning in Higher Education. No. 30, pp. 1-43.

Oksuzler, O. (2008) Does Education Pay off in Turkey? An Ordered Logit Approach, MPRA Paper No. 14375, posted 1.

Ozcan, Y. Z, Ozcan, M. K. and Ucdogruk, S., (2003) Wage Differences by Gender, Wage and Self Employment in Urban Turkey, Journal of Economic Corporation. 24, p.1-24

Sawkins, W.J., Seaman T.P., and Wiliams C. S. (1997) Church Attendance in Great Britain: An Ordered Logit Approach, Applied Economics, 29: 2. February 1.