Hacettepe Journal of Mathematics and Statistics
Volume 41 (4) (2012), 573 – 585

TRANSFORMATION OF A CONTINUOUS TO A DISCRETE VARIABLE: AN APPLICATION OF THE ORDERED LOGIT MODEL TO EXAMINE EFFECTS OF EDUCATION ON INCOME IN TURKEY

Özlem Kiren Gürler*, H. Hatice Özkoç†‡ and Şenay Üçdoğruk*

Received 17 : 11 : 2009 : Accepted 08 : 12 : 2011

Abstract

The aim of this study is to put forward income differences for Turkey according to the ordered logit and human capital models. For this reason, data of the Budget Survey between 2002 and 2006 by TurkStat were used. The dependent variable was the annual yearly disposable income from the main job acquired by an individual, who brought income to the household from his main job. Interpolation was used in categorizing income, a continuous variable, and curve fitting was performed with the least squares approach. In the study, the probabilities in income groups and the changes in probabilities were calculated by the ordered logit model. Later on, it was aimed to put forward the factors that determine income differences for Turkey by using the information on education, age, occupation and marital status of individuals and the social and economic information for a household through the Mincer type of human capital model.

Keywords: Income difference, Continuous variable, Curve fitting, Discrete variable, Ordered Logit, Selectivity bias, Least squares method.

2000 AMS Classification: 62 J 12.

*Dokuz Eylul University, Department of Econometrics, Dokuz Cesmeler 35160, Buca, İzmir, Turkey.
E-mail: (Ö. K. Gürler) ozlem.kiren@deu.edu.tr (Ş. Üçdoğruk) ozlem.kiren@deu.edu.tr

†Mugla University, Faculty of Science, Department of Statistics, 48000 Kotekli, Mugla, Turkey. E-mail: hatice.ozkoc@mu.edu.tr

‡Corresponding Author.

1. Introduction

The distribution of incomes of individuals has constituted one of the important issues of the literature on economics since the 1960s. Income differences, which are quite dwelled upon by public opinion, particularly during periods of crisis, are regarded as one of the essential indicators that determine the development levels of countries. The opinion that the individual is in the classical sense a capital like physical capital, gained a corporate integrity following World War II rather under the influence of technological developments and the studies on the formation of the theory of human capital were accelerated [16]. Besides the physical capital opportunities of countries, the level of sources of human capital is regarded as one of the important elements that affect the development levels of those countries and, therefore, their competition power in the international arena [3]. In order to provide a balance of income in the country, sufficient importance should be attached to human capital alongside physical capital. The role of the individual in the process of production also provides human beings with the quality of capital just like the effect of physical capital on production. In theory, human capital has been regarded as the most important determinant for economic growth and income differences per capita. The education level of individuals appears as the most important factor in the essence of the human capital model. The increases in the education levels of individuals enhance labor force efficiency and, therefore, economic growth [10].

In this study, the aim of the research is to present the components making up the differences in income distribution and to determine probabilistically the effects of those components. As in the ordered logit models, at best we can hope to learn the probabilities of falling into various categories of interest. Once we have estimated such probabilities, we can then discuss how to recover various marginal effects (changes in probabilities) of interest. The information obtained from an interpretation of the results obtained by using a continuous income dependent variable and ordinary least squares approaches would not be useful in terms of presenting the change in income according to the independent variable. When an income dependent variable is used sequentially, interpretation of the recreation model results is more beneficial in terms of the definition of the difference in income distribution. In both data dependent variables, ordered logistic regression produced more accurate estimates of the probability of belonging to the dependent category. OLS predicted values were observed for a number of cases in both studies, and the OLS predicted values were not as strongly related to the dependent binary variable as were the logistic estimates. Also, the logistic estimates were aligned more closely with observed probabilities compared to the OLS estimates. If the purpose of the research is estimating probabilities of the outcome event, logistic regression is the better model. Social science researchers should become more familiar with logistic regression methods and begin to use them when modeling the probability of binary outcomes [14].

An income dependent variable is regularly obtained in a data set. Because of the causes mentioned, that variable should be categorized or in other words should be discretizated. Discretization is to divide the range of the continuous attribute into intervals. Every interval is labeled a a discrete value, and then the original data will be mapped to the discrete values. Most real-world applications of classification algorithms contain continuous numeric attributes. When the feature space of the data includes continuous attributes only, or mixed types of attributes (continuous types along with discrete types), it makes the problem of classification vitally difficult. For example, classification methods based on instance-based measures are generally difficult to apply to such data because the the similarity measures defined on discrete values are usually not compatible with the similarity of continuous values. Alternative methodologies such as probabilistic modeling, when applied to continuous data, require an extremely large amount of data.

In addition, poorly discretized attributes prevent classification systems from finding important inductive rules. For example, if the ages between 15 and 25 map into the same interval, it is impossible to generate a rule about the legal age to start military service. Furthermore, poor discretization makes it difficult to distinguish the non-predictive case from poor discretization. In most cases, inaccurate classification caused by poor discretization is likely to be considered as an error originating from the classification method itself. In other words, if the numeric values are poorly discretized, no matter how good our classification systems are, we fail to find some important rules in the databases. Although discretization influences significantly the effectiveness of classification algorithms, not many studies have been done because it usually has been considered a peripheral issue. A simple method, called the equal distance method, is to partition the range between the minimum and maximum values into N intervals of equal width. Another method, called the equal frequency method, chooses the intervals so that each interval contains approximately the same number of training examples; thus, if $N = 10$, each interval would contain approximately 10% of the examples. However, with both of these discretizations, it would be very difficult or almost impossible to learn certain concepts [7].

There are many classic methods to discretize continuous attribute, including the equal width method, the equal frequency method, the statistic test method, the information entropy method and the clustering-based method, etc. [6]. In this study, by contrast, a discretization process based on the functional structure of the data is applied. By this way, a classification believed to represent the continuous income variable ideally is used.

In this study, the ordered logit and human capital models were estimated for Turkey by considering the household budget surveys obtained by TurkStat between 2002 and 2006. The second section of the study deals with the studies in the literature while its third and fourth sections give information on the method and dataset used. The model results are dealt with in the fifth section and the discussion part takes place in the last section.

2. Literature review

In Turkey, empirical studies on the theory of human capital are limited in number. When these studies are examined, it is observed that the first study was performed by Odekon [12]. Data from population studies at Hacettepe University, dated 1968, were used in the study and it was concluded that education and experience, among the variables of human capital model, could explain 33% of the change in income. Smith [15] examined household and labor force income differences in the Baltic States by the help of an ordered logit regression model on the basis of the human capital model. At the end of the study, it was found that experience and education had essential impacts on income. Specifically, the results indicate considerable increases in returns to education, a significant increase in returns to experience, a substantial increase in occupational wage dispersion, and a large shift in ethnic income differentials.

In another study on the relationship between education and income, Öksüzler [13] investigated the effect of education on the earnings of individuals from Turkey by considering the Mincer [11] earnings equation. Income was defined as a function of education, age and gender and estimations were made by the ordered logit method. It was found that education had a important and positive effect on the income level and it was found out that this return was higher for women.

3. Methodology

The coarsening of a variable might produce an ordered and discrete outcome, although the variable could, in principle, have been measured continuously. A typical example is income data which can grouped. This data is called interval data. Interval data is closely related to standard ordered response models. The specific feature of interval data is that all remaining thresholds are known as well [17].

For a single independent variable, the structural model is

$$y_i^* = \alpha + \beta x_i + \epsilon_i. \tag{1}$$

In the ordinary least square estimate of the regression model, β represents the amount of change in the observed value of the dependent variable which is brought about by a unit change in the independent variable. Since coding of the (ordinal level) dependent variable is arbitrary, this value will depend on the particular coding which is chosen. In the probit model, β represents the amount of change in the dependent variable on its (hypothesized) underlying scale which is brought about by a unit change in the independent variable. Because of the ordinal level assumptions of the n-chotomous probit model, this value is independent of the original coding of the dependent variable, but it is, of course, dependent on the units of the estimated underlying scale for the dependent variable. Consequently, it follows that the β's in two analyses are not directly comparable [9]. Moreover, the linear regression of y on x is that the errors are heteroscedastic and are not normal. In general, the results of the linear regression model only correspond to those of the ordinal regression model if the thresholds are all about the same distance apart. When this is not the case, the linear regression model can give very misleading results [8].

In order to put forward income difference in the study, the continuous income variable was divided into groups and made discrete. For this purpose, the ordered logit and human capital model estimations were made for Turkey in general considering The Household Budget Survey, obtained by TurkStat between 2002 and 2006. Brief theoretical information on the methods used are given in the following section before the results of the model estimations.

3.1. Ordered logit and the parallel slopes assumption. When a variable is ordinal its categories can be ranked in ascending order, but the distances between adjacent categories are unknown. Ordinal outcomes are common in the social sciences. McKelvey and Zavoina [9] studied votes for the 1965 Medicare Bill where each member of Congress was rated as against, weakly for, or strongly for the bill. Marcus and Greene [5] analyzed factors affecting the assignment of Navy recruits into jobs that were ordered as medium skilled, highly skilled, and nuclear qualified [8].

The ordered regression model (ORM) can be derived from a measurement model in which a latent variable y^* ranging from $-\infty$ to ∞ is mapped to an observed variable y.

As usual, in equation (1), y^* is unobserved. The observed y is related to y^* according to the measurement model:

$$y = \begin{cases} 0 & \text{if } y^* \leq \tau_1, \\ 1 & \text{if } \tau_1 < y^* < \tau_2, \\ 2 & \text{if } \tau_2 < y^* < \tau_3, \\ \cdots & \cdots\cdots\cdots\cdots\cdots \\ J & \text{if } \tau_{j-1} < y^* < \tau_j. \end{cases}$$

The τ's are called thresholds or cutpoints [5, 8].

The probability of the observed outcome $y_i = 1, 2, \ldots, j$ given x_i is:

$$\begin{aligned}
&\text{Prob}(y = 1 \mid x_i) = \Phi\left[\tau_1 - \alpha - \beta x_i\right], \\
&\text{Prob}(y = 2 \mid x_i) = \Phi\left[\tau_2 - \alpha - \beta x_i\right] - \Phi\left[\tau_1 - \alpha - \beta x_i\right], \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
&\text{Prob}(y = j \mid x_i) = 1 - \Phi\left[\tau_j - \alpha - \beta x_i\right].
\end{aligned}$$

For the ordered logit model, ϵ (the error term) has a logistic distribution with a mean 0 and a variance of $\pi^2/3$. The pdf and cdf are

$$\text{(2a)} \qquad \lambda(\epsilon) = \frac{\exp(\epsilon)}{[1 + \exp(\epsilon)]^2},$$

$$\text{(2b)} \qquad \Phi(\epsilon) = \frac{\exp(\epsilon)}{1 + \exp(\epsilon)}.$$

3.2. The parallel slopes assumption. A critical assumption of the ordered logit and probit models is that the slope coefficient β of equation (1) does not vary according to the categories of the dependent variable. That is to say, the ordered logit and probit models fit a parallel slopes cumulative model. If one has a reason to believe that the parallel slope assumption is not valid, then the model ought to be estimated using the multinomial logit method, notwithstanding the fact that the dependent variable is clearly ordinal [1]. The idea of parallel regressions can be seen by rewriting the model in terms of the cumulative probability that an outcome is less than or equal to j.

$$\text{(3)} \qquad \text{Prob}(y \leq j | x_i) = \Phi[\tau_j - \alpha - \beta x_i].$$

The cumulative probability is the cumulative distribution function Φ evaluated at $\tau_j - \alpha - \beta x$. Since β is the same for all the j's, equation (3) defines a set of binary response models with different intercepts:

$$\text{(4)} \qquad \tau_j - \alpha - \beta x = (\tau_j - \alpha) - \beta x.$$

The model for $y \leq 1$ is

$$\text{Prob}(y \leq 1) = \Phi[(\tau_1 - \alpha) - \beta x], \text{ with intercept } (\tau_1 - \alpha).$$

The model for $y \leq 2$ is

$$\text{Prob}(y \leq 2) = \Phi[(\tau_2 - \alpha) - \beta x], \text{ with intercept } (\tau_2 - \alpha).$$

In this model, the intercept has changed to $(\tau_2 - \alpha)$, but the coefficient for the variable x is unchanged [8].

3.3. Pooled data. An independently pooled cross section is obtained by sampling randomly from a large population at different points in time (usually, but not necessarily, different years). If a random sample is drawn at each time period, pooling the resulting random samples gives us an independently pooled cross section. One reason for using independently pooled cross sections is to increase the sample size. By pooling random samples drawn from the same population, but at different points in time, we can get more precise estimators and test statistics with more power. Pooling is helpful in this regard only insofar as the relationship between the dependent variable and at least some of the independent variables remains constant over time [18]. In the study, the data of the Household Budget Survey, performed regularly since 2002 by TurkStat, were used and a dataset composed of 63131 individuals, who had a job and acquired activity income in the household, was created from the pooled data of 2002-2006. Information on personal characteristics includes occupation held, education, income, marital status, age and region of residence.

4. The data set

In this study, the activity incomes of individuals aged 12+ from their main job were considered as the dependent variable. However, in an ordered logit model, the dependent income variable should be included as a discrete variable in the model. The purpose here was not to obtain coefficient estimations with the ordinary least squares method but to calculate and interpret estimation probabilities for the categories created. Hence, the ordered logit model was utilized.

The continuous income variable was first divided into five sub-groups as 20% quantiles. However, an aggregation was observed between the 20% and 80% sub-groups. Therefore, it was considered that the division of income using percentages would yield misleading results. In addition, when the income variable is divided into three parts as low, moderate and high, it will not be sufficiently beneficial to examine change in income. In order to observe more clearly the effect on the income group at moderate level, particularly where there was agglomeration, the approach of curve fitting for the data was preferred. For a function $f(x)$ and known functions such as $g_1(x), g_2(x), \ldots, g_m(x)$ which are different from each other, the operation of determining an approach function like

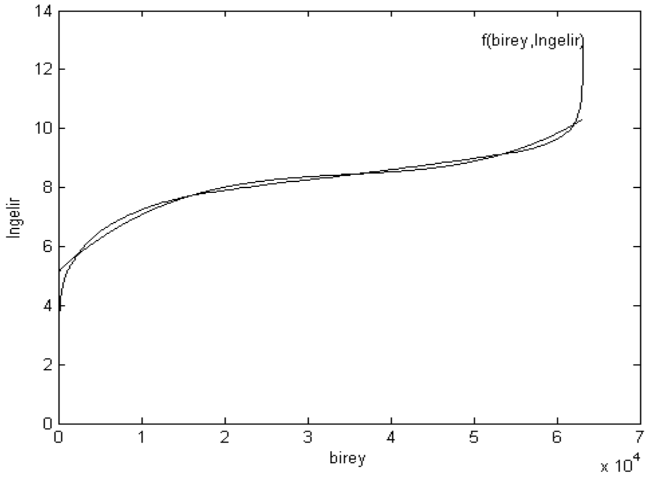
$$g(x) = C_1 g_1(x) + C_2 g_2(x) + \ldots + C_m g_m(x) \tag{5}$$

is called curve fitting. The best criterion for curve fitting is the least squares approach. Curve fitting, which is based on the least squares approach, depends on solving $m + 1$ simultaneous linear equation systems for determining the polynomial coefficients of m^{th} order. The cubic spline approach is also used for curve fitting. Nevertheless, it is not possible to make the necessary grouping for ordered logit since no functions are defined between two points when using a spline [2].

The theoretical income function, regarded as a polynomial of third degree, using the curve fitting approach was found as in Equation 6 by the MATLAB program.

$$f(x) = (1/15428796088736)x^3 + (1/146925306)x^2 + (48/189715)x + (901/175) \tag{6}$$

Figure 1. The Income Curve and Parabola of Third Degree



The cut-off points of the curve, composed of real observations and theoretical values, were used for transforming the continuous variable of income into a discrete variable. For this purpose, the points where the curve based on real observations diverged from the theoretical function were determined. Income was divided into six categories and the continuous variable was transformed into a discrete variable. Table 1 below presents the number of observations in the groups for the income variable which was transformed into a discrete variable.

Studies on personal income distribution in the literature of economics concentrate on income distribution modeling and measurement of income inequality. Personal income distribution is utilized to determine income inequalities in a country at a specific time slot. Personal income distribution can be used to examine the development of income inequality in a country over time [4].

With surveys on personal income distribution, two different goals may be targeted. One of them is to reveal the level of inequality of any income distribution at a specific moment, while the second one is to make comparisons in terms of level of inequality among various income distributions.

Table 1. Number of Observations in the Categories Obtained for Activity Income

Dependent Variable (Classes)	Frequency	Percentage	Lower Limit	Upper Limit
1	2098	3.32	2.1491	284.733
2	13904	22.02	284.8688	2242.622
3	20166	31.94	2243	4748.125
4	17471	27.67	4749.104	8153
5	8430	13.35	8155.75	25123.5
6	1062	1.68	25145.46	378000
Total	63131	100		

In Table 1, it is seen that in terms of income distribution a percentage of low income individuals exceed high income individuals'. More than half of the individuals involved in the sample belong to middle income level and below,

$$(\%3.32 + \%22.02 + \%31.94 = \%57.28).$$

The descriptive statistics of the independent variables used in the ordered logit and human capital models are given in Table 2. The information on education, age, occupation, marital and social security status of 63131 individuals, 211677 of whom obtained activity income, was used in the analyses. Year dummies were created in order to examine whether the increase in income changed by year. The variable of income was deflated on the basis of the year 2003 due to the difference in inflation between 2002 and 2006. Furthermore, corrections were made on the data of 2002-2004 since six zeros were dropped from the Turkish Lira as of 2005, and the income variable was used after taking the logarithm. When Table 2 is examined, it is observed that income was the same on average in Turkey in general. It is understood that 45% of individuals were primary school graduates. It is seen that compared with the past the mean education level of individuals increased after eight-years of education become compulsory. Individuals working without any social security constituted 46%. Moreover, more individuals were working in agriculture than in the other occupational groups.

Table 2. Descriptive Statistics

Variable	Mean	Std. Dev.
Dependent variable: ln(income)	8.1700	1.0988
Activity Income	5770.648	7875.155
Age		
6-24	0.1392	0.3461
25-39	0.4383	0.4962
40+	0.4225	0.4940
Education		
Education (Continuous)	7.4873	3.8903
Illiterate	0.0416	0.1997
Literate but completed no school	0.0417	0.1998
Primary school graduate	0.4516	0.4977
Junior high school graduate	0.1308	0.3372
High school graduate	0.2158	0.4114
Vocational school graduate	0.0315	0.1746
Graduate of University and higher	0.0870	0.2818
Social Security		
Social Security Organization and Bank	0.3085	0.4619
Bağkur (Social Security Agency for artisans and the self-employed)	0.1158	0.3200
Retirement fund	0.1102	0.3131
Unregistered	0.4654	0.4988
Occupation		
Legislators, senior officials and managers	0.1124	0.3158
Professionals	0.0687	0.2529
Technicians and associate professionals	0.0550	0.2281
Clerks	0.0569	0.2316
Service workers and shop and market sales workers	0.1139	0.3177
Skilled agricultural and fishery workers	0.1563	0.3632
Craftsmen and workers in related jobs	0.1909	0.3930
Plant and machine operators and assemblers	0.1043	0.3057
Year Dummies		
2002	0.1618	0.3682
2003	0.4082	0.4915
2004	0.1388	0.3458
2005	0.1444	0.3515
2006	0.1468	0.3539
Marital Status		
Married	0.7974	0.4020
Single	0.2026	0.4020
Area		
Urban	0.6586	0.4742
Rural	0.3414	0.4742
Gender		
Male	0.8222	0.3824
Female	0.1778	0.3824
N	63131	

5. Model results

In this study, the ordered logit model estimation of income groups was estimated for Turkey during the period 2002-2006. Two methods were used in the interpretations of coefficients.

i) Partial change in y^*,
ii) Discrete change in probabilities for the dummy variables.

When the results of the goodness of fit of the model are examined before interpreting the model parameters, the null hypothesis of the LR test statistic is that all of the regression coefficients are simultaneously equal to zero. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. A small p-value from the LR test, 0.000, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero.

Table 3 presents the result of the ordered logit model obtained for the pooled dataset between 2002 and 2006 for Turkey. When Table 3 is examined and the standardized coefficients for y are considered, it is observed that men have higher incomes than women (a standard deviation of 0.4059). It can be stated that the incomes of individuals aged 25–39 and 40 + increased more than those aged below 25 (standard deviations of 0.3753 and 0.5163, respectively)[§].

Table 3. Ordered Logit Model Estimation Results

Dependent Var.: ln(Income)*	$\widehat{\beta}^{**}$	Std. Err.	z	bStdX	bStdY	bStdXY	SDofx
Male	0.9854	0.0215	45.797	0.3768	0.4059	0.1552	0.3824
Age							
25-39	0.9111	0.0293	31.095	0.4521	0.3753	0.1862	0.4962
40+	1.2535	0.0317	39.512	0.6192	0.5163	0.2550	0.4940
Education							
Literate	0.1457	0.0529	2.752	0.0291	0.060	0.0120	0.1998
Primary school graduate	0.6175	0.0403	15.315	0.3074	0.2544	0.1266	0.4977
Junior high school graduate	0.7733	0.0454	17.027	0.2608	0.3185	0.1074	0.3372
High school graduate	1.1365	0.0449	25.260	0.4676	0.4681	0.1926	0.4114
Vocational school graduate	1.4944	0.0614	24.311	0.2609	0.6156	0.1075	0.1746
University Graduate and above	2.1149	0.0569	37.133	0.5961	0.8712	0.2455	0.2818
Social Security							
Social Security Organization and bank	1.1505	0.0201	57.220	0.5314	0.4739	0.2189	0.4619
Bağkur (Soc. Security Agent. – artisans and self-employed)	1.8962	0.029	64.612	0.6069	0.7811	0.2500	0.3200
Retirement Fund	1.7595	0.0326	53.851	0.5510	0.7248	0.2269	0.3131

[§] No interpretations of standard deviation will be made for the remaining coefficients in order to avoid repetition.

Table 3. (Continued)

Occupation							
Legislators, senior officials and managers	1.9447	0.0352	55.196	0.6142	0.8011	0.2530	0.3158
Professionals	0.9664	0.0479	20.171	0.2444	0.3981	0.1007	0.2529
Technicians and associate professionals	1.1474	0.0406	28.223	02617	0.4726	0.1078	0.2281
Clerks	0.7834	0.0396	20.052	0.1814	0.3227	0.0747	0.2316
Service workers and shop and market sales workers	0.4303	0.0298	14.401	0.1367	0.1773	0.0563	0.3177
Skilled workers in agriculture, animal husbandry, hunting and fishery products	0.6558	0.0305	21.499	0.2382	0.2701	0.0981	0.3632
Craftsmen and workers in related jobs	0.5967	0.0262	22.693	0.2345	0.2458	0.0966	0.3930
Plant and machine operators and assemblers	0.8500	0.0307	27.666	0.2598	0.3501	0.1070	0.3057
Urban	0.1313	0.0181	7.232	0.0623	0.0541	0.0256	0.4742
Year Dummies							
2002	-0.336	0.0221	-15.147	-0.1238	-0.1384	-0.0510	0.3682
2004	0.251	0.0242	8.873	0.0745	0.0888	0.0307	0.3458
2005	0.247	0.0228	10.834	0.0870	0.1019	0.0358	0.3515
2006	0.251	0.0226	11.069	0.0889	0.1034	0.0366	0.3539
Married	0.520	0.0251	20.665	0.2092	0.2143	0.0861	0.4020
_cut1	0.2501	0.0514					
_cut2	3.0684	0.0512					
_cut3	5.1011	0.0540					
_cut4	7.1736	0.0568					
_cut5	9.8907	0.0652					
-2LLR (26)	35038.9	Prob = 0.000					N= 63131

NOTE: *: Income is composed of six ordered groups.

**: All coefficients are statistically significant at the significance level of 1%.

Base categories:
Female, illiterate, those who do not have any social security, workers in jobs that do not require any qualifications as well as single residents in rural areas.

bStdX= X standardized coefficient,
bStdY= y standardized coefficient,
bStdXY= completely standardized coefficient,
SDofX= standard deviation of X.

In addition, when year dummies are examined, it can be noted that there was an increase in income in 2003 when compared to 2002. In the following years, the level of increase remained almost the same, in other words, no income differences occurred. When education is examined, income increased from those who were literate to university graduates. Nevertheless, the highest increase in income took place in the occupational group of legislators and senior officials when compared to the employees working in unqualified jobs. This was followed by the occupational groups of associate professionals and the occupational groups of plant and machine operators, respectively.

Table 4. Variations in Estimated Probabilities

Variable	*	Variat.	1	2	3	4	5	6
Gender								
Male	0 → 1	0.0706	-0.0164	-0.1511	-0.0444	0.1583	0.0495	0.0039
Age								
25-39	0 → 1	0.0721	-0.0105	-0.1130	-0.0928	0.1486	0.0625	0.0053
40+	0 → 1	0.0985	-0.0142	-0.1521	-0.1293	0.1976	0.0902	0.0078
Education								
Literate	0 → 1	0.0117	-0.0016	-0.0178	-0.0157	0.0244	0.0099	0.0008
Primary school graduate	0 → 1	0.0490	-0.0071	-0.0777	-0.0622	0.1026	0.0410	0.0344
Junior high school graduate	0 → 1	0.0633	-0.0071	-0.0822	-0.2006	0.1209	0.0634	0.0055
High school graduate	0 → 1	0.0919	-0.0102	-0.1185	-0.1471	0.1701	0.0970	0.0087
Vocational school graduate	0 → 1	0.1172	-0.0096	-0.1198	-0.2221	0.1665	0.1679	0.0172
University Graduate and above	0 → 1	0.1556	-0.0125	-0.1545	-0.2998	0.1706	0.2649	0.0313
Social Security								
Social Security Organization and bank	0 → 1	0.0923	-0.0115	-0.1287	-0.1367	0.1781	0.0908	0.0080
Bağkur (Social Security Agency – artisans and self-employed)	0 → 1	0.1446	-0.0125	-0.1522	-0.2690	0.1917	0.2184	0.0236
Retirement Fund	0 → 1	0.1360	-0.0119	0.0144	-0.2515	0.1905	0.1969	0.0207
Occupation								
Legislators, senior officials and managers	0 → 1	0.1473	-0.0126	-0.1536	-0.2757	0.1895	0.2272	0.0250
Professionals	0 → 1	0.0789	-0.0078	-0.0939	-0.1352	0.1408	0.0880	0.0080
Technicians and associate professionals	0 → 1	0.0929	-0.0086	-0.1046	-0.1656	0.1557	0.1124	0.0106
Clerks	0 → 1	0.0634	-0.0067	-0.0797	-0.1065	0.1195	0.0674	0.0060
Service workers and shop and market sales workers	0 → 1	0.0359	-0.0043	-0.0494	-0.0514	0.0706	0.0318	0.0027
Skilled workers in agriculture, animal husbandry, hunting and fishery products	0 → 1	0.0535	-0.0063	-0.0725	-0.0818	0.1051	0.0511	0.0044
Craftsmen and workers in related jobs	0 → 1	0.0486	-0.0059	-0.0725	-0.0719	0.0969	0.0450	0.0038
Plant and machine operators and assemblers	0 → 1	0.0692	-0.0074	-0.0872	-0.1141	0.1297	0.0726	0.0064
Urban	0 → 1	0.0103	-0.0015	-0.0170	-0.125	0.0221	0.0083	0.0006
Year Dummies								
2002	0 → 1	0.0259	0.0044	0.0462	0.0272	-0.0565	-0.0196	-0.0016
2004	0 → 1	0.0173	-0.0023	-0.0262	-0.0235	0.0360	0.0148	0.0012
2005	0 → 1	0.0199	-0.0026	-0.0299	-0.0273	0.0413	0.0171	0.0014
2006	0 → 1	0.0202	-0.0027	-0.0303	-0.0277	0.0419	0.0174	0.0014
Marital Status								
Married	0 → 1	0.0369	-0.0072	-0.0732	-0.0384	0.0869	0.0295	0.0024

*: 0 → 1 shows the change from 0 to 1.

Base categories:
Female, illiterate, those who do not have any social security, workers in jobs that do not require any qualifications as well as single residents in rural areas.

By making use of the coefficient estimations obtained from ordered logit model and the averages of independent variables, it is possible to find out the marginal impacts of the probabilities of the income groups. If dummy variables are present in the independent variables, it is misleading to interpret marginal impacts [8]. In this case, it is more suitable to use the measurement of discrete change.

With this in mind, the results for the model yielding variations in probabilities are given in Table 4. When Table 4 is examined, it can be stated that there is lower probability for men to be included in a low-income group than for women, and that the probability of men to be included in an upper income group is higher towards the high income groups.

For each income group, the probability of being included in an upper income group increases with increasing education level. However, this is not applicable in high income groups. Amazingly, the probability of being a primary school graduate turned out to be higher in the group concerned in comparison to junior high school, high school and vocational school.

Another striking point is that the probability of being in the first and sixth income groups is rather lower than the probabilities of being in other income groups. When the dependent variable of income is examined, it is observed that there are quite extreme values in both groups. When the first income group is considered, it is observed that there are individuals with almost no income while the sixth income group includes those with astronomic income levels. The mean income of 2098 individuals in the first income group was 160 TL, whereas the mean income of 1062 individuals in the sixth income group was 46000 TL. It is thought-provoking that the number of individuals in the first income group was approximately twice those in the sixth income group and that the sum of individuals included in both groups constituted 5% of all observations. Lowering these values, indications of the income gap among individuals in the society, to reasonable levels should form the foundation of economic policies. Particularly reducing these individuals to below one in a hundred of the society is imperative for a reduction of the income gap.

When variations by year are examined, it is observed that the probability of being in the upper income groups was low in 2002 in comparison to 2003, while the probability of being in the high income groups increased in post-2003 years in comparison to 2003 but that this increase remained almost the same between 2004 and 2006, in other words, there was no change in income.

6. Conclusion

In this study, it was intended to examine income differences with the data obtained from Household Budget Surveys for the 2002–2006 period by TurkStat. When the results of the ordered logit model performed with a pooled dataset were considered, it was observed that the effect of gender on income is still maintained in Turkey. This means that men earn a higher income than women.

A striking point in the results of the ordered logit model is that income generally increased by year. Even though the effect of implemented economic programs and the fall in inflation on income is pleasing, it is thought-provoking that the rate of increase in income has recently decreased.

Another striking point is that the probabilities of being in the first and sixth income groups remained rather lower than the other income groups. These extreme levels, which include individuals with very low and very high incomes, can in a sense be regarded as an indication of the income gap in the society.

The fact that the individual has social security has an important impact on the increase in income.

Acknowledgements

The authors thank the referees for their valuable comments which improved the paper.

References

[1] Borooah, V. K. *Logit and Probit–Ordered and Multinomial Models* (Sage University Paper **138**, California, 2002).

[2] Chapra, S. C. and Canale, R. P. *Mühendisler için Sayısal Yöntemler* (Literatür Yayınları (In Turkish), Ankara, 2008).

[3] Erdoğan, S. *Temel İnsan Sermayesi Modeli : Seçilmiş İllerde Ekonometrik Yaklaşım* (In Turkish), D. E. Ü. İ. İ. B. F. Dergisi, İzmir **14** (1), 75–95, 1999.

[4] Goodman, A., Johnson, P. and Webb, S. *Inequality in the UK* (Oxford University Press, Oxford, UK, 1997).

[5] Greene, W. *Econometrics Analysis* (Prentice Hall, New Jersey, 2003).

[6] Jiang, S., Li, X. and Zheng, Q. *Approximate Equal Frequency Discretization Method*, Intelligent Systems, GCSI'09. ERI, 514–518, 2009.

[7] Lee, C. *Discretizing continuous attributes using information theory*, Computer and Information Sciences **3733**, 493–502, 2005.

[8] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables* (Sage Publications, California, 1997).

[9] McKelvey, R. D. and Zavoina, W. *A statistical model for the analysis of ordinal dependent variable*, The Journal of Mathematical Sociology **4**, 103–120, 1975.

[10] Metin, K. and Üçdoğruk, Ş. *İstanbul ilinde gelir farklılıklarını belirleyen etmenler:İnsan Sermayesi Modeli (1994)*, Ekonomik Yaklaşım **8** (27) (In Turkish), 283–302, 1997.

[11] Mincer, J. *Schooling and Earnings* (Columbia University Press, New York, 1974).

[12] Odekon, M. *The Impact of Education on The Size Distribution of Earnings in Turkey* (Unpublished Ph.D. Dissertation, State University of New York, Albany, 1977).

[13] Öksüzler, O. *Eğitim ve gelir ilişkisi: Türkiye Örneği* (In Turkish), TÜİK 16. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, 291–300, 2007.

[14] Pohlmann, J. T. and Leitner, D. W. *A comparison of ordinary least squares and logistic regression*, Ohio J. Sci, **103** (5), 118–125, 2003.

[15] Smith, K. *Determinants of household and labor income in the Baltic States: Soviet and post-Soviet results*, The European Journal of Comparative Economics **4** (1), 3–24, 2007.

[16] Tunç, M. *Kalkınmada İnsan Sermayesi Yaklaşımları ve Türkiye'de İnsan Sermayesi Boyutunun Analizi* (Unpublished PhD Thesis (In Turkish), DEÜ, Sosyal Bilimler Enstitüsü, İzmir, 1997).

[17] Winkelmann, R. and Boes, S. *Analysis of Microdata* (Springer-Verlag, Berlin, Heidelberg, 2006), 200–201.

[18] Wooldridge, M. J. *Introductory Econometrics:A Modern Approach* (5th Edition, Prentice Hall, Ohio, 2003).