Ordinal Logit and Multilevel Ordinal Logit Models: An Application on Wealth Index MICS-Survey Data

Muhammad Arfan College of Statistical and Actuarial Sciences University of the Punjab, Lahore muhammadarfan25@gmail.com

Rehan Ahmad Khan Sherwani College of Statistical and Actuarial Sciences University of the Punjab, Lahore rehan.stat@pu.edu.pk

Abstract

Ordinal logistic regression models are used to predict the dependent variable, when dependent variable is of ordinal type in both the situation for single level and multilevel. The most used model for ordinal regression is the Proportional Odd (PO) model which assumes that the effect of the each predictor remains same for each category of the response variable. To estimate the wealth index of household in the province Punjab the proportional odds model is used. The wealth index is an order categorical dependent variable having five categories. The data MICS (2014), a multiple indicator cluster survey conduct by Punjab bureau of statistics was used in this article. The data was recorded at different level such as individual level (household level), district level and division level. The secondary data MICS contains a sample of 41413 household collected from both rural and urban areas of the province Punjab. In the present study analysis were made for single level (household level) and two levels (division level). After fitting the proportional odds model for the single level the proportionality assumption is tested by the brand test whose results suggest that all the predictors fulfill assumption of proportional odds. The significance value suggests that all the predictors have significant effect on the wealth index. The variation due to division level was estimated by two level ordinal logistic regression equal to 5.842, and the Intra Class Correlation ICC is equal to 0.6397 which show that 63.97% of total variation is due to division level.

Keywords: Discrete Choice Models, Ordinal Categories, Ordinal Logit, Multilevel Ordinal Logit.

1. Introduction

Discrete choice models is a class of model in which the response variable Y takes the counted values such as 0, 1, 2, and so on to a finite number of values (Joe, 2008). In the case where the natural ordering exist in the dependent variable. For example the grade of a high school student may, very good, good, satisfactory, poor, and very poor. Another example opinion about a product of soap may be strongly opposed, opposed, neutral, support and strongly support for such a categorically we can code the 1 for" strongly opposed", 2 for "opposed", 3 for "neutral", 4 for " support" and 5 for "strongly support". Here the values are not quantitative but a natural ordering exist between the values. It is not necessary that the difference between the category 2,3 is same as the difference between the category 4 and 5. Commonly, for the prediction of an ordinal response variable ordinal logistic regression is used (Bello et al., 2016; Christensen, 2010).

1.1. Ordinal Logistic Regression Model

Many models for the ordinal dependent variable has been studied in literature; the basic objective of ordered logit models is calculation of accumulative probability for dependent variable being greater than the j^{th} category (Peterson and Harrel, 1990; Brant, 1990; Agresti, 1996; Liu and Agresti, 2005). McCullagh (1980) called this model as the proportional odds model (POM), where the basic assumption is that effect of independent variable is same across all the categories of dependent variable, this is also called the proportional odds assumption or parallel lines assumption. To relax this proportional odds assumption we used some generalized model that relax the proportional odds assumption for some predictors or for all the predictors, called as the partial proportional odds model (PPOM) by Peterson and Harrell (1990).

For a response variable Y with C categories and a set of predictors X having the effect parameters β the probability of response variable being less than or equal to category j can be modeled by the logistic distribution as

$$\gamma_{j} = \Pr(\mathbf{Y} \le y_{j} | X)$$

$$\Pr(\mathbf{Y} \le y_{j} | X) = \frac{\exp[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})]}{1 + \exp[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})]}$$

$$i = 1, 2, 3, \dots, C - 1$$

Where $j = 1, 2, 3, \dots, C-1$

The above proportional odds model gives the cumulative probability γ_i of category j and for the response variable having categories C we find the C-1 cumulative probabilities as for the last category the cumulative probability is always equal to one. The above model can also be written as

$$\Pr(\mathbf{Y} \le y_j | \mathbf{X}) = \frac{1}{1 + \exp\left[-\alpha_j + \left(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}\right)\right]}$$

The probability that response variable lies in the category greater j is

$$\Pr(Y > y_{j}|X) = 1 - \Pr(Y \le y_{j}|X)$$

$$\Pr(Y > y_{j}|X) = 1 - \frac{\exp[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})]}{1 + \exp[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})]}$$

$$\Pr(Y > y_{j}|X) = \frac{1}{1 + \exp[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})]}$$

The odds of response variable being less than or equal to category j to category greater than j can be found as

$$\frac{\Pr(Y \le y_j | X)}{\Pr(Y > y_j | X)} = \exp[\alpha_j - (\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik})]$$

And the logit model is the natural log of odds ratio and is the linear function of k independent variables

$$Log\left[\frac{\Pr\left(Y \le y_{j} | X\right)}{\Pr\left(Y > y_{j} | X\right)}\right] = \alpha_{j} - \left(\beta_{1} X_{i1} + \beta_{2} X_{i2} + \dots + \beta_{k} X_{ik}\right)$$

Where $j = 1, 2, 3, \dots, C-1$

The proportional odds model assumes that the explanatory variables have the same on the response variable across all the categories of the response variable, this is called the proportional odds assumption. Under the assumption of proportional odds the β_s remains same and only intercepts varies for different categories of response variable. As the sign of β_s is negative (subtracted) they show how the one unit increases in predicter increase in log odds of being in the category greater than j. The cumulative logit model

$$\gamma_{j} = \frac{\exp\left[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})\right]}{1 + \exp\left[\alpha_{j} - (\beta_{1}X_{i1} + \beta_{2}X_{i2} + \dots + \beta_{k}X_{ik})\right]}$$

 α_j are the intercepts and are different for each comparison ordinal categorical variable, and relation between α_j is $\alpha_1 < \alpha_2 < \dots < \alpha_{c-1}$ to ensure that $\gamma_1 < \gamma_2 < \dots < \gamma_{c-1}$.

The slope coefficients $\beta_1, \beta_2, \dots, \beta_k$ are the same for all the categories of dependent variable, for continuous variables the slope coefficients change in log odds for one unit change in predictor and for nominal predictors the slope coefficient represent the effect of each category of nominal variable as compared to reference category.

1.2. Multilevel Ordinal Logistic Regression Model

Many types of data, including observational study collected from the human and biological sciences, are nested in clustered. For example, children belong to same parents are more to same in their mental and physical look than children selected at random from a large population. Individual further may be nested within localities and institution such as schools etc. Multilevel structure can also be exist in time involving studies, where individual's responses may be correlated over time (Chan et al., 2015). Multilevel models are for hierarchical nested data structure by allowing error components at different level of hierarchy. For example, a two level used for child outcomes where child outcomes are nested within schools. Multilevel model estimate the residual at both level, students and schools. Thus the total residual variance is divided into parts one for between schools (level two units, schools residual) and one for within schools (between level one units, students residuals).

In the field of science, education and related field mostly research application has the dependent variable of ordinal nature and is non normal nature in their distribution. For example the marks of students in an examination is categorized as (1: Extra ordinary, 2:

Good, 3: Satisfactory, 4: Average, 5: Fail) here the dependent variable is measured on ordinal scale having five ordered categories, or proficiency scores obtained by mastery testing process, such as when we are interested to find out the factors affecting students level of proficiency in reading or English ("e.g., 1: Below basic; 2: Basic; 3: Proficient; 4: Beyond proficient"). The regression analysis for ordinal categorical data, non ordinal data, rates, proportions and for all other types of non normal outcomes, time involving data are most of the part of models that are generally belongs to generalized linear models. For continuous data the ordinary least square regression models are known as the special case of generalized linear models. We also used the generalized linear models to show the pattern of most variables which are of limited or discrete type and also are of non normal type. The multilevel analysis is used to estimate the model and to estimate the variation is response occur within cluster and between clusters, the units of higher level. The variation between clusters represent random effect due to inclusion of cluster level in the model. In multilevel modeling this variation is most important and interesting part of the analysis.

Multilevel ordinal logistic regression model is the analysis of hierarchical and ordinal dependent variable. it is used to model the ordinal categorical dependent by one or more independent variables. The ordinal categorical dependent follow the logistic distribution and nested with higher levels (O'Connell, 2010; Khiari and Rejeb, 2015).

Let the Y_{cij} ordered categorical response of i^{th} individual in the j^{th} cluster with C ordered categories coded as $C = 1, 2, \dots, c$. Then the cumulative probability for ordered response up to category c is $P_{ijc} = \Pr(Y_{ij} \le c)$. The multilevel random intercept (base model) cumulative log odds model for ordinal response is written as

$$log\left(\frac{\Pr\left(Y_{ij} \le c | X\right)}{\Pr\left(Y_{ij} > c | X\right)}\right) = \alpha_{c} + u_{oj}$$

Which measure of odds of Y_{cij} being in the category less than or equal to C as compared to greater than the category C, and u_{oj} is the random effect of level two units and is assumed to follow normal distribution $N(0, \tau_0)$. The above model random intercept model when there is no explanatory variables. When model also have some fixed explanatory variables then the above model can be written as

$$log\left(\frac{\Pr\left(Y_{ij} \le y_c | X\right)}{\Pr\left(Y_{ij} > y_c | X\right)}\right) = \alpha_c + X_{ij}\beta + u_{oj}$$

Where X_{ij} is the data matrix of fixed predictors, hence fixed effect β s are the same as for simple proportional odds model.

The sample size required to find reliable estimate of multilevel models for ordinal data depends on many factors, for example the complication of the model, estimate of cluster level variance, method of parameters estimation etc. Some guidelines about sample size are provided from recent simulation studies for multilevel order logit model. Austin (2010) used a random intercept logistic regression model, whereas Moineddin *et al.* (2007) used both random intercept and random slopes logit model in which both the intercept and the slope randomly vary across clusters. For the model with random intercept case, the estimates are comparatively good with every estimation methods even for 10 to 15 clusters, with cluster size the average equal to 10. If the size of clusters are small then more cluster are required to obtain reliable estimates. For the random slope model more clusters with larger size are required say 30 clusters with average size of 30.

Intra class correlation coefficient is used as assessment of how many variation in the response categories lies at the level two (group level). When logistic model is used the residual at level one are assumed to follow the standard logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$ =3.29. ICC the within groups variation for dichotomous and ordinal

outcomes (Snijder and Bosker 2008) is defined as

$$ICC = \frac{\tau_0}{\tau_0 + 3.29}$$

where τ_0 is variance level two error term and 3.29 is variance of standard logistic distribution.

2. Data and Methodology

In present study for the analysis and measure of Wealth Index Quintile, we used Multiple Indicator Cluster Data (MICS-2014). The target population of the Survey is all the area of province Punjab rural and urban which are defined by population census of 1998 and changing made by Government thereafter. The MICS survey conducted by Bureau of statistics in the 2004, used multistage stratified sampling to select the a sample of house hold, a complete list of enumeration area has been taking from the Punjab bureau of statistics. The sampling frame was first divided into 9 Division and thirty six administrative districts in the Punjab, and the divided into rural and urban areas of the province Punjab. So first stage sampling consist of selection of enumeration area and second stage sampling consist of selection of sample of house hose.

The value of all physical, natural and financial assets owned by a household is called the wealth. The wealth index is a cumulative measure of a household living standard. It is a composite index includes of all the asset of ownership variables, wealth index indicate the proxy wealth of household level. Wealth index is calculated using the data about the ownership of household assets, such as material used in the construction of house, types of access to water for drinking, facilities of sanitation, bicycles, television, cars etc. To

Muhammad Arfan, Rehan Ahmad Khan Sherwani

construct the wealth index of a house we need all indicator that that are used as assets. Wealth index vary from household to household, locality to locality and country to county. After 1990s, wealth indices are become major tool for measuring the economic status of a household. These are considered effective indicator for Scio- economic position, standard of living, health standard of household (Córdova, 2009).

Wealth characteristic of household has a greater effect on the health. Wealth index can be used for identification the problems, specially for poor household, as poor has unequal access to health care facilities compared to wealthy. DHS program that is partially funded by the world bank developed the wealth index, that is used by governments to identify whether education, health services, and other essential are reaching to poorest households. Wealth index is used to check how economics status of household affects health education etc (Garenne and Hohmann, 2003; Hong et al., 2006).

Estimation of wealth index is based on the data collected by household questionnaire, which include the questions related to ownership of number of televisions, dwelling status, characteristics related to building of house, source of drinking water, number of cars, toilet facility and other characteristics related to wealth status (Howe et al., 2008, Van Campenhout, 2007). Assign the weights or factor score to each household assets which is collected by principal components analysis. Standardized the assets scores by the standard normal distribution that had zero mean and unit standard deviation. Then break points these standardized scores to obtain the wealth index quintile as "Lowest", "Second", "Third", "Fourth" and "Highest". Standardized score are assigns to each assets of household, and these score differs for each asset depending on that household own or not that asset. The scores of all assets are summed and household are ranked on the basis of total score. A single index for asset is identified by the data about the entire country sample and used for all households. These indices are same for rural and urban living households and no separate index is fined for rural and urban areas (Rasbash et al., 2002). The present research is planned to measure the wealth index quintile for house hold in the province Punjab by ordinal logistic regression model and by multilevel ordinal logistic regression model.

3. Results and Discussion

Descriptive statistical analysis is most important part of the study and is the best way to check the nature of the data. Following table shows the descriptive statistics for household data.

Variables	Categories	Wealth Index Quintile						
		Lowest Second Third			Fourth Highest			
		Count (%)	Count (%)	Count (%)	Count (%)	Count (%)		
Area of	Rural	6358 (28.3)	6083 (27.1)	5217 (23.2)	3427 (15.2)	1384 (6.2)		
nousenoid	Urban	519 (4)	1140 (9.8)	2239 (17.3)	3896 (30.1)	5133 (39.7)		
Dwelling Status	Own	6758 (20.6)	6922 (21.1)	6885 (21)	6378 (19.5)	5820 (16.4)		
Status	Rented	119 (4.5)	301 (11.4)	571 (21.7)	945 (35.9)	697 (26.5)		
Education	Non	4796 (34.5)	3545 (25.5)	2727 (19.6)	2001 (14.4)	814 (5.9)		
of Household	Primary	1107 (17.9)	1538 (24.9)	1487 (24.1)	1277 (20.7)	767 (12.4)		
11000001010	Middle	507 (11)	959 (20.8)	1182 (25.6)	1153 (25)	814 (17.6)		
	Secondary	392 (5.9)	940 (14.3)	1508 (22.9)	1911 (29)	1842 (27.9)		
	Higher	75 (1.8)	241 (5.8)	252 (13.4)	981 (23.8)	2280 (55.2)		
Source of	Pipe Into Dwelling	75 (2.2)	368 (10.9)	699 (20.65)	955 (28.2)	1291 (38.1)		
Drinking	Public Tab	278 (19.2)	385 (26.6)	378 (26.1)	207 (19.8)	122 (8.4)		
water	Tube well	254 (11.1)	265 (11.6)	394 (17.2)	584 (25.6)	788 (34.5)		
	Hand Pump	5490 (50.9)	3043 (28.2)	1507 (14)	594 (5.5)	146 (1.4)		
	Motorized Pump	737 (4.6)	3087 (19.5)	4287 (27)	4441 (28)	3308 (20.9)		
	Cart With Small Tank	43 (3)	74 (5.1)	190 (13.1)	457 (31.6)	683 (47.2)		
	Bottled Water	0 (0)	0 (0)	1 (0.5)	5 (2.7)	179 (98.8)		
Household	Yes	5617 (34.7)	4615 (28.5)	3528 (21.8)	1790 (11)	650 (4)		
Own Any Animal	No	1260 (6.6)	2606 (13.6)	3928 (20.5)	5533 (28.8)	5867 (30.6)		
Electricity	Yes	5438 (16)	7206 (21.2)	7450 (22)	7323 (21.6)	6516 (19.2)		
	No	1439 (98.4)	17 (1.2)	6 (0.4)	0 (0)	1 (0.1)		
Type of Fuel Used for Cocking	Electricity	18 (12.4)	33 (22.8)	43 (29.7)	30 (20.7)	21 (14.5)		
	LPG	0 (0)	41 (2.3)	217 (12.1)	786 (43.8)	751 (41.9)		
	Natural Gas	21 (0.2)	348 (2.9)	1767 (14.8)	4292 (36)	5487 (46.1)		
	Wood	3440 (29)	3702 (30.1)	3388 (27.6)	1565 (12.7)	188 (1.5)		
	Animal Dung	1545 (29.9)	1804 (35)	1291 (25)	464 (9)	56 (1.1)		
	Agricultural Crop Residue	1853 (45.2)	1295 (31.8)	750 (18.3)	186 (4.5)	14 (0.3)		

 Table 3.1: Descriptive Analysis for the Province Punjab

Results presented in table 3.1 depicted that for rural area 28.3% household have lowest, 27.1% have second, 23.2% have third and 6.2% have highest wealth index quintile. Which show that in rural area mostly household have low wealth index quintile and small percentage of household have greater wealth index quintile. For the urban area 4% have lowest, 9.8% have second, 17.3% have third, 30.1% have fourth and 39.7% have highest wealth index quintile that show in urban area mostly people have greater wealth index quintile. For the second explanatory variable household own dwelling the descriptive statistics result show that distribution of household have own house is lowest are 20.6%, second are 21.1%, third are 21%, fourth are 19.5% and highest are 16.4%. For the household which lived in rented house 4.5% have lowest, 11.4% have second, 21.7% have third, 35.9% have fourth and 26.5% have highest wealth index quintile. Most of the household having greater wealth index lived in the rented house. For the variable education of household head (HH), for illiterate HH 34.4% have lowest, 25.5% have second, 19.6% have third, 14.4% have fourth and 5.9%, so the HH with non education mostly have lower wealth index quintile. The HH having higher education 1.8 % have

lowest, 5.8% have second, 13.8% have third, 23.8% have fourth and 55.2% have highest wealth index quintile, so results show that if the HH is more educated he will be likely to have greater wealth index quintile. The independent variable household own any animal for the category the Yes 34.7% lowest, 28.5% second , 21.8% third, 11% fourth and 4% have highest wealth index quintile and for the category No distribution is as 6.6% have lowest, 13.6% have second, 20.5% have third, 28.5% have fourth and 30.6% have highest wealth index quintile. so mostly household have any animal have the lower wealth index quintile with no animals have greater wealth index quintile. For the next predictor household have electricity facility the household that have electricity facility 16% have lowest, 21.2% have second, 22% have third, 21.6% have fourth and 19.2% have highest and that have not electricity facility 98.4% have lowest, 1.2% have second, 0.4% have third, 0% have fourth and 0.1% have highest wealth index quintile. Household with no electricity facility belong to lower wealth electricity. The next predictor the main source of drinking water show that household having pipe into 2.2% belong to lowest, 10.9% belong to second, 20.65% belong to third, 28.2% belong to fourth and 38.1 belong to highest. Households that have hand pump 50.9 belong to lowest maximum are at lowest wealth index quintile. For bottled water 98.8% belong to highest wealth index quintile. The household which are using electricity for cocking are 12.4% belong to lowest, 22.8% belong to second, 29.7% belong to third, 20.7% belong to fourth and 14.5% belong to highest wealth index quintile. Most of the household that used LPG for cocking belong to higher wealth index quintile 43.8% have fourth and 41.9% have highest. Household that are using animal dung and agricultural crop residue have lower wealth index quintile.

3.1. Ordinal Logistic Regression Models

The present study is to estimate the wealth index quintile a ordered categorical response variable Y_{ij} having five ordered categories (Lowest, Second, Third, Fourth, Highest). Ordered logit model estimate the cumulative probability γ_j or cumulative log odds

 $log(\frac{\gamma_j}{1-\gamma_j})$ up to j^{th} category where j = 1, 2, 3, 4. One category (Last) of the response

variable is taken as reference category and cumulative probability for reference category is always equal to one. The cumulative logit probability model takes the form as:

$$\begin{split} \gamma_{j} &= \frac{1}{1 + \exp[-\alpha_{j} + \beta_{ll}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9}]} \\ \log[\frac{P(Y \leq j)}{P(Y > j)}] &= \log(\frac{\gamma_{j}}{1 - \gamma_{j}}) \\ \log(\frac{\gamma_{j}}{1 - \gamma_{j}}) &= \alpha_{j} - (\beta_{1l}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9}) \\ \log(\frac{\gamma_{1}}{1 - \gamma_{1}}) &= \alpha_{1} - (\beta_{1l}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9}) \\ \log(\frac{\gamma_{2}}{1 - \gamma_{2}}) &= \alpha_{2} - (\beta_{1l}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9}) \end{split}$$

$$\log(\frac{\gamma_{3}}{1-\gamma_{3}}) = \alpha_{3} - (\beta_{1l}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9})$$

$$\log(\frac{\gamma_{4}}{1-\gamma_{4}}) = \alpha_{4} - (\beta_{1l}x_{1} + \beta_{2l}x_{2} + \beta_{3l}x_{3} + \beta_{4l}x_{4} + \beta_{5l}x_{5} + \beta_{6l}x_{6} + \beta_{7l}x_{7} + \beta_{8}x_{8} + \beta_{9}x_{9})$$

Where $log(\frac{\gamma_1}{1+\gamma_2})$ is the log odds of wealth index quintile lowest to second, third, fourth and highest, $\log(\frac{\gamma_2}{1+\gamma_2})$ is the log odds of lowest and second to third, fourth and highest, $\log(\frac{\gamma_3}{1+\gamma_2})$ is the logit of lowest, Second and third to fourth and highest and $\log(\frac{\gamma_4}{1+\gamma_2})$ is the log odds of wealth index quintile lowest through fourth to highest. $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the intercepts for ordered logit model called cut points one for each comparison. $\beta_{1/x_1} = 1$ varies over 1 to 2 the categories of x_1 (area of household), first category urban is taken as reference. $\beta_{2l}x_2$ = where l varies over 1 to 5 the categories of x_2 (education of household head) more over first category non/preschool is taken as reference. $\beta_{3l}x_3$ where l varies over 1 to 7 the categories of x_3 (main source of drinking water) and first category pipe into dwelling is taken is reference. $\beta_{4l}x_4$, l varies over 1 to 2 the categories of x_4 (dwelling status of household) and first category own of x_4 is taken as reference. $\beta_{5l}x_5 = 1$ varies over 1 and 2 the categories of x_5 (household own any animal) more over first category yes is taken as reference. $\beta_{6l}x_6$ where 1 varies over 1,2 the categories of x_6 (electricity facility) and first category yes is taken as reference. $\beta_{7l}x_7$ where l varies over 1 to 6 the categories of x_7 (type of fuel used for cocking) more over the first category electricity if taken as reference. In $\beta_8 x_8$, β_8 is the slope coefficient for x_8 number of household member. In $\beta_9 x_9$, β_9 is the slope coefficient for x_9 (total children aged 1-17 year).

Variables	Parameter Estimates			95% confidence		
Variable option	В	S.D	Exp(B)	Sign.	9.006	97.807
Cut Points						
$\alpha_1 (\leq Lowest)$	-2.159	0.115			-2.494	-1.824
α_2 (<=Second)	0.029	1.029			-0.303	0.363
α_3 (<=Third)	2.198	9.006			1.963	2.532
α_4 (<= fourth)	4.562	95.774			4.244	4.920
x_1 (Area of House Hold)						
Rural (ref)	0		1			
β_{12} (Urban)	0.349	0.028	1.418	0.000	0.294	0.405

 Table 3.2: Parameters Estimates for Proportional Odds Model.

Muhammad Arfan, Rehan Ahmad Khan Sherwani

Variables	Parameter Estimates				95% confidence	
x_2 (Education of HH)						
None/ Pre School (ref)	0		1			
β_{22} (Primary)	0.525	0.030	1.690	0.000	0.465	0.584
β_{23} (Middle)	0.917	0.034	2.502	0.000	0.851	0.983
β_{24} (Secondary)	1.444	0.031	4.236	0.000	1.383	1.504
β_{25} (Higher)	2.340	0.041	10.284	0.000	2.259	2.420
x_3 (Source of drinking water)						
Pipe Into Dwelling (ref)	0		1			
β_{32} (Public Tab)	-1.205	0.062	0.299	0.000	-1.327	-1.0836
$eta_{_{33}}$ (Tube Well)	-0.289	0.054	0.749	0.000	-0.396	-0.182
β_{34} (Hand Pump)	-2.243	0.044	0.106	0.000	-2.329	-2.158
β_{35} (Motorized Pump)	-0.089	0.038	0.914	0.019	-0.164	-0.014
$eta_{ m 36}$ (Cart With Small Tank)	0.182	0.065	1.199	0.005	0.055	0.308
$eta_{ m 37}$ (Bottled Water)	2.919	0.436	18.527	0.000	2.059	3.778
X_4 (Dwelling Status)						
Own (ref)	0		1			
β_{42} (Rented)	-0.377	0.041	0.686	0.000	-0.457	-0.297
x_5 (HH Own any Animals)						
Yes (ref)	0		1			
β_{52} (No)	1.041	0.026	2.832	0.000	0.990	1.091
x_6 (Electricity)						
Yes (ref)	0		1			
eta_{62} (No)	-4.581	0.215	0.010	0.000	-5.003	-4.159
x_7 (Type Of Fuel Used For Cooking)						
Electricity (ref)						
β_{72} (LPG)	1.480	0.169	4.396	0.000	1.148	1.812
$eta_{_{73}}$ (Natural Gas)	1.624	0.164	5.071	0.000	1.302	1.945
eta_{74} (Wood)	-1.249	0.163	0.207	0.000	-1.568	-0.929
eta_{75} (Animal Dung)	-1.326	0.164	0.264	0.000	-1.651	-1.006
eta_{76} (Agricultural Crop Residue)	-1.973	0.165	0.138	0.00	-2.298	-1.648
x_8 (Number of HH Member)						
β_8	0.228	0.005	1.256	0.000	0.217	0.239
x_9 (Total children Aged 1-17 Year)						
β_9	-0.263	0.008	0.786	0.000	-0.278	-0.248

-2loglikelihood =70224.39; $\chi^2_{(21)}$ = 43628.3, p < 0.001; *Pseudo* R^2 = 0.3832

Statistical results presented in table 3.2 showed that p-values for each predictors is less than 0.01 except one motorized pump, so all the predictors with all the categories are statistically significant. Exp(B) is odd ratio for the each predictor variables, the interpretations are made on the basis of above results. The four intercept are used to differentiate the category of wealth index quintile one for each comparison. These are also called is the cut points of comparison -2.159 is used for comparison of lowest to second, third, fourth and highest, 0.029 is used to compare category lowest, second to category third, fourth and highest, 0.182 is used as intercept of comparison of category lowest, second and third to category fourth and highest and 4.562 is used as intercept to compare the category lowest, second, third and fourth to category highest of wealth index quintile. The household living in the urban area OR (exp(b)) = 1.418, show that that odds of being in the greater category of wealth index quintile are 1.418 time more as compare to household living in urban area. So the house hold live in urban area have greater wealth index quintile. Parameter estimate for the variable education of education of household head HH having primary education have 1.690 time more odds of being in grater category as compare to those who are illiterate, household head having middle education are 2.502 time more likely to be in the grater category as compare to illiterate, household head having secondary education have 4.236 time more odd of being in the grater category of wealth index quintile compared to illiterate and household head that have higher education are 10.284 time more likely to be in greater category of wealth index quintile as compared to those how are illiterate. It can be interpreted as that a household having higher education is more likely to be in greater category of wealth index quintile. For the nominal independent variable main source of drinking water household having public tab have 0.299 time less odds of being in greater category compared to household having pipe into dwelling, household using tube well and hand pump as source of drinking water have 0.749 and 0.106 time respectively less odds of being in greater category of wealth index quintile, household having motorized pump and cart with small tank of water have 0.914 and 1.199 time more odds of being in higher category of response variable and household using bottled water as drinking water has 18.522 time more odds of being in greater category of wealth index quintile as compared to household having piped into dwelling. Parameter estimate for the variable dwelling status of household for rented house household odds of being in greater category of wealth index quintile decrease by the factor 0.686 as compared to those having their own house. For the variable household have any animal the parameter estimate show that household with no animal have 2.832 time more odds of being in grater category of wealth index quintile as compared to those having any animal. For the predictor household have electricity facility, the odd of being in the greater category of wealth index quintile for household that have not electricity facility decreased by the factor 0.010 as compared to those how have electricity facility. For the independent variable type of fuel used for cocking, household using LPG have 4.396 time more odds of being in the greater category of wealth index quintile as compared to household using electricity for cocking, household using natural gas for cocking have 5.071 time more odds of being in the greater category of wealth index quintile, for household using wood odds of being in higher category decreased by the factor 0.207 as compared to those how are using electricity, household using animal dung have odds 0.264 time less for being in higher category of wealth index quintile as compared to electricity user household and for the household using agricultural crop residue as fuel for cocking have 0.138 time less odds of being in higher category of wealth index quintile as compared electricity used for

Muhammad Arfan, Rehan Ahmad Khan Sherwani

cocking. For the explanatory variable number of household members the odds of being in higher category of wealth index quintile increased 1.256 time for each household member. For the variable number of children aged 1-17 the parameter estimate show that odds of being in greater category of wealth index quintile decreased by the factor 0.786 for every increased child aged 1-17.

The value of -2 log likelihood for the estimated model is 70224.39 and the value of chisquare (43628.3, df=21, p<0.001) showed the statistically significance of the model. Pseudo R- Square in logistic regression is one minus ratio 2 log likelihood of final model to 2 log likelihood of null model it is not equal to R^2 in multiple linear regression but in tell us that approximately how many variation in dependent variable is explained by independent variables. A greater value of pseudo R-Square show better fit. For the present model pseudo R- Square show that approximate 38.32% vitiation in the wealth index quintile is explained by all the independent variables included in the model.

3.2. Multilevel Ordinal Logistic Regression Models

In the previous section we have fit the ordinal logit model for household level (single level) and explained the parameter estimates for the household level. As households are nested in the high level of hierarchy such as households are nested in division (level 2) so we need to fit a two level multilevel ordinal logit model to estimate the response variable wealth index quintile. In two level random intercept model we allow the cut points to vary at different level of hierarchy. We take these random variation as a part of total random variation, i.e. we distribute the whole random variation into parts associated with different levels of hierarchies. In the present problem the estimation of wealth index quintile we use two level multilevel random intercept ordinal logit model to find variation due division level of household.

A two level random intercept ordinal logit model to estimate the wealth index quintile with fixed predictors the cumulative log odds model can by written as

$$\log(\frac{\gamma^{(c)}_{ij}}{1-\gamma^{(c)}_{ij}}) = \alpha^{(c)} + u_{0j} + \beta_{1k}x_1 \dots \text{mod}_{ijk} + \beta_{2k}x_2 \dots \text{mod}_{ijk} + \beta_{3k}x_3 \dots \text{mod}_{ijk} + \beta_{4k}x_4 \dots \text{mod}_{ijk} + \beta_{5k}x_5 \dots \text{mod}_{ijk} + \beta_{6k}x_6 \dots \text{mod}_{ijk} + \beta_{7k}x_7 \dots \text{mod}_{ijk} + \beta_{8}x_8 \dots \text{mod}_{ij} + \beta_{9}x_9 \dots \text{mod}_{ij}$$

where c = 1, 2, 3, 4, $\gamma^{(c)}_{ij}$ represent the cumulative probability of i^{th} household in j^{th} division less than or equal to category C where C = 1, 2, 3, 4. u_{oj} is random variation of j^{th} division and is assumed to follow $N(0, \sigma^2_{u_{oj}})$.

Log odds of category lowest to all above

$$\log(\frac{\gamma^{(1)}_{ij}}{1-\gamma^{(1)}_{ij}}) = \alpha^{(1)} + u_{0j} + \beta_{1k}x_1 \dots \text{mod}_{ijk} + \beta_{2k}x_2 \dots \text{mod}_{ijk} + \beta_{3k}x_3 \dots \text{mod}_{ijk} + \beta_{4k}x_4 \dots \text{mod}_{ijk} + \beta_{5k}x_5 \dots \text{mod}_{ijk} + \beta_{6k}x_6 \dots \text{mod}_{ijk} + \beta_{7k}x_7 \dots \text{mod}_{ijk} + \beta_8x_8 \dots \text{mod}_{ij} + \beta_9x_9 \dots \text{mod}_{ij}$$

Log odds of lowest and second to all above categories

$$\log(\frac{\gamma^{(2)}_{ij}}{1-\gamma^{(2)}_{ij}}) = \alpha^{(2)} + u_{0j} + \beta_{1k}x_1 \dots \text{mod}_{ijk} + \beta_{2k}x_2 \dots \text{mod}_{ijk} + \beta_{3k}x_3 \dots \text{mod}_{ijk} + \beta_{4k}x_4 \dots \text{mod}_{ijk} + \beta_{5k}x_5 \dots \text{mod}_{ijk} + \beta_{6k}x_6 \dots \text{mod}_{ijk} + \beta_{7k}x_7 \dots \text{mod}_{ijk} + \beta_8x_8 \dots \text{mod}_{ij} + \beta_9x_9 \dots \text{mod}_{ij}$$

Log odds for the categories lowest-third to category fourth and second

$$\log(\frac{\gamma^{(3)}_{ij}}{1-\gamma^{(3)}_{ij}}) = \alpha^{(3)} + u_{0j} + \beta_{1k}x_1 \cdot \text{mod}_{ijk} + \beta_{2k}x_2 \cdot \text{mod}_{ijk} + \beta_{3k}x_3 \cdot \text{mod}_{ijk} + \beta_{4k}x_4 \cdot \text{mod}_{ijk} + \beta_{5k}x_5 \cdot \text{mod}_{ijk} + \beta_{6k}x_6 \cdot \text{mod}_{ijk} + \beta_{7k}x_7 \cdot \text{mod}_{ijk} + \beta_8x_8 \cdot \text{mod}_{ij} + \beta_9x_9 \cdot \text{mod}_{ij}$$

Log odds of category lowest-fourth to category highest

$$\log(\frac{\gamma^{(4)}_{ij}}{1-\gamma^{(4)}_{ij}}) = \alpha^{(4)} + u_{0j} + \beta_{1k}x_1 \dots \text{mod}_{ijk} + \beta_{2k}x_2 \dots \text{mod}_{ijk} + \beta_{3k}x_3 \dots \text{mod}_{ijk} + \beta_{4k}x_4 \dots \text{mod}_{ijk} + \beta_{5k}x_5 \dots \text{mod}_{ijk} + \beta_{6k}x_6 \dots \text{mod}_{ijk} + \beta_{7k}x_7 \dots \text{mod}_{ijk} + \beta_8x_8 \dots \text{mod}_{ij} + \beta_9x_9 \dots \text{mod}_{ij}$$

C = 1, 2, 3, 4, predictor are same as used in the single level analysis.

Table 3.3: Parameter estimates of the random intercept ordinal logit with fixed predictors.

Variables	Estimates		
Fixed	В	S.D	Exp(B)
Cut Points			
$\alpha^{(1)}$ (<= Lowest)	-4.119	0.273	0.0162
$\alpha^{(2)}$ (<=Second)	-0.466	0.271	0.6275
$\alpha^{(3)}$ (<=Third)	3.091	0.272	21.999
$\alpha^{(4)}$ (<= fourth)	7.014	0.176	1112.094
$x_1 \cdot \text{mod}_{ij}$ (Area of House Hold)			
Rural (ref)	0		1
β_{12} (Urban)	-0.446	0.045	0.6401
$x_2 . \text{mod}_{ij}$ (Education of HH)			
None/ Pre School (ref)	0		1
β_{22} (Primary)	-0.839	0.060	0.4321
β_{23} (Middle)	-1.486	0.066	0.2262
β_{24} (Secondary)	-2.344	0.061	0.0959
β_{25} (Higher)	-3.784	0.080	0.0227
$x_3 . \mathrm{mod}_{ij}$ (Source of drinking water)			
Pipe Into Dwelling (ref)	0		1

Variables	Estimates		
β_{32} (Public Tab)	1.822	0.102	6.1842
β_{33} (Tube Well)	0.436	0.093	1.5465
β_{34} (Hand Pump)	3.324	0.077	27.7712
β_{35} (Motorized Pump)	0.099	0.068	1.1040
β_{36} (Cart With Small Tank)	-0.319	0.114	0.7268
β_{37} (Bottled Water)	-3.094	0.939	0.0453
$x_4 . \mathrm{mod}_{ij}$ (Dwelling Status)			
Own (ref)	0		1
β_{42} (Rented)	0.611	0.079	1.8422
$x_5 . \mathrm{mod}_{ij}$ (HH Own any Animals)			
Yes (ref)	0		1
β_{52} (No)	-1.667	0.049	0.1888
$x_6 \cdot \text{mod}_{ij}$ (Electricity)			
Yes (ref)	0		1
β_{62} (No)	5.583	0.518	265.868
$x_7 . \text{mod}_{ij}$ (Type Of Fuel Used For Cooking)			
Electricity (ref)	0		1
β_{72} (LPG)	-2.316	0.269	0.0986
β_{73} (Natural Gas)	-2.483	0.259	0.0834
β_{74} (Wood)	2.155	0.257	8.6278
β_{75} (Animal Dung)	2.539	0.260	12.667
β_{76} (Agricultural Crop Residue)	3.320	0.264	27.6603
$x_8 \cdot \text{mod}_{ij}$ (Number of HH Member)			
β_8	-0.282	0.009	0.7542
x_9 . mod _{ij} (Number of Children aged 1-17 year)			
β_9	0.360	0.013	1.4333
Var. Component	Variance	$e(\tau_0) =$	5.842 (0.094)
ICC	$\overline{\tau_0}$	$\frac{\tau_0}{+3.29}$	= 0.6397

For multilevel model the estimation method MQL 1^{st} order produced can produced the biased estimated there we used the PQL method for estimation of parameter of multilevel models. The result of MLWIN, PQL 2^{nd} order estimation are shown in the table above. Where the value of Intra Class Correlation (ICC) 0.6397 show that 63.97% variation in dependent variable "wealth index" in at division level, (level two). By using the random intercept model with fixed predictors the variation at division level is 5.842. Intercept (cut point) to find the odds of lowest wealth index to all above categories is -4.119, cut point for odds of category lowest, second to third fourth and highest wealth index is -

0.466, intercept for odds of category lowest, second and third to category fourth and highest is 3.091 and the cut point for the odds of category lowest, second , third and fourth to highest is 7.014. cut points follow the relationship, -4.119 < -0.466 < 3.091 < 7.014.

4. Conclusion

In present research, multilevel ordinal logistic regression were fitted to measure the wealth index by using fixed effect and random effect. Fixed effect level 1, model was fitted the results of level 1 (household level) were same as the results of ordinal logistic regression. For two level ordinal logistic regression, only intercept model the variability at division level is 11.554 and Intra Class Correlation ICC is 0.7783 which show that 77.83% of total variability lies at division level. We conclude that mostly variation lies at level two (division level). Next we fit the two level multilevel random intercept model for dependent variable "wealth index quintile" and a set of fixed predictors which are " area of household", education of household head", "type of fuel used for cocking", "main source of drinking water", " number of household members", "total children aged 1-17 year", "household own any animals", "electricity facility", "dwelling status of household". The odds ratios for two level random intercept model with fixed predictors are almost same as for the level one ordinal logistic regression model and are interpreted in the same manner. For multilevel model the more important thing the variation at higher level, for our model the random variation at level two (Division level) is Variance $(\tau_0) = 5.842$ and Intra Class Correlation ICC = 0.6397. Intra class correlation show that the 63.97% of variation lies at division level. The level one residual has a standard logistic distribution with variance $\frac{\pi^2}{3}$ = 3.29. The much of total variation lies at level two units.

References

- 1. Agresti, A. (1996). An Introduction to Categorical Data Analysis. New York: Wiley & S Sons, ISBN 0-471-11338-7.
- 2. Austin, J.T., Yaffee, R.A., Hinkle, D.E. (1992). Logistic regression for research in higher education. In: Higher Education: Handbook of Theory and Research, Paulsen, M. (ed.). Springer, New York, p379–410.
- 3. Bello, A. O., Oguntolu, F. A., Adetutu, O. M., & Ojedokun, J. P. (2016). Application of Bootstrap Re-sampling Method to a Categorical Data of HIV/AIDS Spread across different Social-Economic Classes. *arXiv preprint arXiv:1609.07806*.
- 4. Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 46 (4), 1171–1178.
- 5. Chan, M. T., Yu, D., & Yau, K. K. (2015). Multilevel cumulative logistic regression model with random effects: Application to British social attitudes panel survey data. *Computational Statistics & Data Analysis*, 88, 173-186.

- 6. Christensen, R.H.B. (2010) Ordinal Regression Models for Ordinal Data, sofware and manual downloadable from: <u>http://cran.r-</u> project.org/web/packages/ordinal/index.html.
- 7. Córdova, A. (2009). Methodological note: Measuring relative wealth using household asset indicators. *Insights Series No. 10806*.
- 8. Garenne, M., & Hohmann-Garenne, S. (2003). A wealth index to screen high-risk families: application to Morocco. *Journal of Health, Population and Nutrition*, 235-242.
- 9. Hong, R., Banta, J. E., & Betancourt, J. A. (2006). Relationship between household wealth inequality and chronic childhood under-nutrition in Bangladesh. *International Journal for Equity in Health*, 5(1), 15.
- 10. Howe, L. D., Hargreaves, J. R., & Huttly, S. R. (2008). Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerging themes in epidemiology*, *5*(1), 3.
- 11. Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics & Data Analysis 52, 5066–5074.
- 12. Khiari, M., & ben Rejeb, J. (2015). Determination of the Regional Impact on Innovation with an Ordinal Logit and a Multilevel Analysis. *Procedia-Social and Behavioral Sciences*, 195, 592-602.
- 13. Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, *14*(1), 1-73.
- 14. McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109-142.
- 15. Moineddin, R., Matheson, F.I. & Glazier, R.H. (2007). A simulation study of sample size for multilevel logistic regression models. BMC Medical Research Methodology 7:34.
- 16. O'Connell, A. A. (2010). An illustration of multilevel models for ordinal response data. In *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8).*
- 17. Peterson, B., & Harrell F.E. (1990). Partial Proportional Odds Models for Ordinal Response Variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 39, No. 2 (1990), pp. 205-217.
- 18. Rasbash, J., Steele, F., Browne, W.J., & Prosser, B. (2002). A Users Guide to MLwiN. Version 2.0. Centre for Multilevel Modelling, University of Bristol, Bristol.
- Snijders, T.A.B., & Berkhof, J. (2008). Diagnostic Checks for Multilevel Models. In Handbook of Multilevel Analysis (ed. de Leeuw J and Meijer E) pp. 141–175. Springer, New York.
- 20. The Multiple Indicator Cluster Survey (MICS) Punjab, 2014.
- 21. Van Campenhout, B. F. (2007). Locally adapted poverty indicators derived from participatory wealth rankings: A case of four villages in rural Tanzania. *Journal of African Economies*, *16*(3), 406-438.