Linear Possibility Model for Ordered Categorical Data: A Similar Way of Analysis to Regression Analysis

Amin I. Adam (Associate Professor in Applied Statistics, Dept. of Statistics, Faculty of Economics and Political Sciences,)
Omdurman Islamic University, Omdurman, P. O. Box 382, Sudan

Abstract

This research is concerned with introducing the linear possibility model for ordered categorical data, a model which is extended from the idea of the probit model. In linear possibility model the response variable for ordered categorical data is having a wider range of values, and consequently, the conditional expected values for this response given a number of regressors will be outside the range of 0-1. This research shows how the problems surrounding the linear possibility model can be solved, to a large extent, allowing the model to give simple and straight forward interpretations for the relationships between the categorical variables similar to regression analysis. The study gives empirical estimates of the effects between the variables and explains estimates corresponding to the ordering nature of the categorical variables under concerned. The application data for this research are collected from a random sample of students at the Omdurman Islamic University. The ordered response categorical variable for the study is the academic performance of students, which is assumed to be associated with three categorical variables: the specialization of the students, whether the students live with their families or not, and the educational level of their guardians. The result showed that the conditional possibility of the academic performance of the students is lower by almost a third for all the students whose specialization is social science. Likewise, the conditional possibility of the academic performance of the students is higher by almost a third for all the students whose the educational level of their guardian is secondary. AS for those, whose specialization is natural science and the educational level of their guardian is primary or lower, the conditional possibility of their academic performance is 3.791, which is "good". The data of the study is analysed with the aid of SPSS, (Statistical Package for the Social Sciences) and Minitab.

Key Words: **Probit model, Linear models, Regression analysis, Categorical data**.

النموذج الراجح الخطي للبيانات الفئوية المرتبة: تحليل ومسار مشابه لتحليل الانحدار

ملخص الدراسة

يعنى هذا البحث بالتعريف بالنموذج الراجح الخطي لتحليل البيانات الفئوية المرتبة، وهو النموذج الذي تم تعميمه من فكرة النموذج الاحتمالي. في النموذج الراجح الخطي للبيانات الفئوية المرتبة يكون للمتغير التابع قيماً متعددة ومن ثم قيماً متوقعة مشروطة ذات مدى أوسع خارج نطاق الصفر والواحد بالطبع. أوضح هذا البحث كيف أن المشاكل التي يمكن أن تحيط بالنموذج الراجح الخطي يمكن معالجتها، بشكل كبير، مما يتيح للنموذج لإعطاء تفسيرات سهلة وسلسة للعلاقة بين المتغيرات الفئوية كتلك التي يعطيها نموذج الانحدار، وقد أعطت الدراسة تقديرات لأثر العلاقة بين المتغيرات محل الدراسة كما أوضحت التفسيرات المقابلة لطبيعة كون هذه المتغيرات الفئوية مرتبة. وكانت البيانات قد جمعت من عينة عشوائية من طلاب جامعة أم درمان الإسلامية، وكان مستوى التحصيل الأكاديمي بمثابة المتغير الفئوي التابع في الدراسة، ويرتبط بثلاثة من المتغيرات الفئوية؛ مساق الطلاب، إذا كان الطلاب يعيشون مع أسرهم أم لا، والمستوى التعليمي لأولياء أمورهم. أوضحت الدراسة أن القيمة الراجحة الشرطية لأداء الطلاب الأكاديمي تكون أدنى بحوالي الثلث للطلاب الذين يكون تخصصهم أدبياً، وبالمثل، أوضحت الدراسة أن القيمة الراجحة الشرطية لأداء الطلاب الأكاديمي تكون أعلى بحوالي الثلث للطلاب الذين يكون المستوى التعليمي لأولياء أمورهم ثانويا (مستوى المرحلة الثانوية)، أما للطلاب الذين يكون تخصصهم علميا، ويكون المستوى التعليمي لأولياء أمورهم أوليا أو أقل، فإن القيمة الراجحة الشرطية لأدائهم الأكاديمي تكون 3.791، أي "جيداً". هذا، وقد تم استخدام الحزمة الإحصائية للعلوم الاجتماعية (SPSS)، وحزمة (Minitab) للمساعدة في تحليل البيانات.

1- Introduction:-

In the probit model the response variable is binary, having values as zeros and ones. Therefore, the conditional expected values for the response variable are theoretically expected not to lie outside the range of 0-1, and hence being as probability expected values. In linear possibility model, however, the response variable for ordered categorical data is having a wider range of values, and, consequently, the conditional expected values for this response given a number of regressors will be outside the range of 0-1.

For a set of categorical variables, we consider a response ordered categorical variable Y to be dependent on a number of k other categorical variables, ordered or otherwise. We assume the response variable to have categories being ordered (ascendingly) as 1,2,...,c. For the c_1 categories of the first categorical repressor variable, c_2 categories of the second categorical regressor variable, ..., and c_k categories of the kth categorical regressor variable, we can write a linear possibility model for Y as

$$Y_i = \beta_0 + \beta_{11}X1_i^1 + \beta_{12}X1_i^2 + ... + \beta_{1(c_1-1)}X1_i^{c_1-1} + \beta_{21}X2_i^1 + \beta_{22}X2_i^2 + ... + \beta_{2(c_2-1)}X2_i^{c_2-1} ... + \beta_{k1}Xk_i^1 + \beta_{k2}Xk_i^2 + ... + \beta_{k(c_k-1)}Xk_i^{c_k-1} + u_i, \quad i = 1,2,...,n \qquad (1)$$

Where

$Y_i = 1,2,...,c$ for the ordered categories of the response categorical variable

$$X1_i^1 = \begin{cases} 1, & \text{for the first category of the first categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$X1_i^2 = \begin{cases} 1, & \text{for the second category of the first categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$X1_i^{c_1-1} = \begin{cases} 1, & \text{for the } (c_1-1) \text{ category of the first categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$X2_i^1 = \begin{cases} 1, & \text{for the first category of the second categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$X2_i^2 = \begin{cases} 1, & \text{for the second category of the second categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$X2_i^{c_2-1} = \begin{cases} 1, & \text{for the } (c_2-1) \text{ category of the second categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$Xk_i^1 = \begin{cases} 1, & \text{for the first category of the kth categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$Xk_i^2 = \begin{cases} 1, & \text{for the second category of the kth categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

$$Xk_i^{c_k-1} = \begin{cases} 1, & \text{for the } (c_k - 1) \text{ category of the kth categorical regressor variable} \\ 0, & \text{otherwise} \end{cases}$$

u_i = the error term

In the above model, the conditional expectation of Y given the k categorical variables can be interpreted as the conditional possibility[1] that the specific event of Y will occur. The parameter β_{11} measures the differential effect[2] for the first category of the first categorical variable, compared with otherwise. Similarly, β_{12} stands for the differential effect for the second

category of the first categorical, compared with otherwise. All other parameters give the corresponding differential effects for the specific categories of the categorical variables. The parameter β_0 represents the expected possibility of Y for absence of all the differential effects of the above specified categories of the categorical variables. We can choose any other combination of categories of the regressed categorical variables to be represented by β_0. The differential effects would then be measured from that combination. However, the numerical variables values of the conditional means will be the same regardless of the starting position.

For three categorical regressor variables, in particular, with the first two variables being binary (1,0) and the third one being ordinal having four categories (1,2,3, and 4), we can have a complete set of all the conditional
means for equation (1) as in table(1).

Table(1): The Expected Values Of Y For Models (1) with Three Assumed **Categorical Regressor Variables.**

X1	X2	X3	For Model (1)
1	1	1	$\beta_0+\beta_{11}+\beta_{21}+\beta_{31}$
1	1	2	$\beta_0+\beta_{11}+\beta_{21}+\beta_{32}$
1	1	3	$\beta_0+\beta_{11}+\beta_{21}+\beta_{33}$
1	1	4	$\beta_0+\beta_{11}+\beta_{21}$
1	0	1	$\beta_0+\beta_{11}+\beta_{31}$
1	0	2	$\beta_0+\beta_{11}+\beta_{32}$
1	0	3	$\beta_0+\beta_{11}+\beta_{33}$
1	0	4	$\beta_0+\beta_{11}$
0	1	1	$\beta_0+\beta_{21}+\beta_{31}$
0	1	2	$\beta_0+\beta_{21}+\beta_{32}$
0	1	3	$\beta_0+\beta_{21}+\beta_{33}$
0	1	4	$\beta_0+\beta_{21}$

[1] The name `linear possibility model' is given for such a situation where the conditional expectation of the dependent variable gives the possible outcome value for the dependent variable and not the probability that specific event occurs.

[2]The term differential effect is used because the specified parameter gives the effect due to the difference between the existence and non-existence of the corresponding variable o r interaction.

0	0	1	$\beta_0+\beta_{31}$
0	0	2	$\beta_0+\beta_{32}$
0	0	3	$\beta_0+\beta_{33}$
0	0	4	β_0

2- The LPM & The Ordinary Least Squares (OLS) Method

The form of the LPM in equation (1) can be written in a matrix form as

$$\mathbf{Y = X\ \beta = U} \qquad \mathbf{(2)}$$

where **Y** represents the column vector of n observations on the response variable Y, **β** is the column vector of k unknown parameters, **X** gives the nxk matrix of observations of the categorical regressor variables (X1, X2, ..., Xk.) with the first column of 1's representing the intercept term, and **U**

accommodates the vector of the disturbances u_i. Model (1) then looks like the usual regression models and hence can be estimated by the OLS method. The familiar OLS formula for the unknown **β**'s is

$$\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}\ X'Y} \qquad \mathbf{(3)}$$

and their variance-covariance matrix **S** is given by

$$\mathbf{S = (X'X)^{-1}\sigma^2} \qquad \mathbf{(4)}$$

with the variance of the error σ^2 (which is also equals the variance of **Y**) to be estimated by

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \qquad \mathbf{(5)}$$

where e_i is the sample residual and k is the number of parameters in the model (including the constant). The significance of each β_i coefficient can therefore be tested by using

$$t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \qquad \mathbf{(6)}$$

which follows the t-distribution with n-k degrees of freedom, with $se(\beta_i)$ being the standard error of β_i (the square root of the corresponding diagonal element in the variance-covariance matrix). For the overall significance of the model, this is given by

$$\mathbf{F} = \frac{\mathbf{(\hat{\beta}'X'Y)}/(k-1)}{\mathbf{(Y'Y - \hat{\beta}'X'Y)}/(n-k)} \qquad (7)$$

which follows the F-distribution with k-1 and n-k degrees of freedom. The t and F-distributions, however, are only approximate here since the dependent variable is discrete. A commonly used measure of goodness-of-fit of the

model is R^2, the coefficient of determination which is given by

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2} \qquad \mathbf{(8)}$$

and the previous F-test, which can be rewritten as

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad (9)$$

serves as a test for R^2 being significantly different from zero.

3- The LPM & The Weighted Least Squares (WLS) Method

It is simple to apply the OLS method to models (1) but obviously this will be surrounded by some technical and logical problems. These special problems are: nonnormality of the error term u_i, heteroscedasticity, and the possibility of the estimated response outcome lying outside the bounds (1 to c for our response Y).

It is true that the OLS does not require the error term u_i to be normally distributed, but it is implicitly assumed so for the purpose of statistical inference, that is, hypothesis testing, prediction, etc. Obviously, the assumption of normality for u_i is hard to attain for an LPM because like the response variable, u_i takes on only discrete values. So, it cannot be assumed to be normally distributed; in fact it follows a multinomial distribution. However, the violation of the normality assumption may not be as serious as it looks because we know that the OLS point estimates still remain unbiased (Johnston and DiNardo (2001)). Furthermore, and based on the central limit theorem, as the sample size increases, the OLS estimators tend to be normally distributed generally (Gujarati(2004)).

The second problem with the error terms u_i is that their variances are no longer equal, i.e., heteroscedastic, even though $E(u_i)=0$ and $E(u_iu_j)=0$ for $i \neq j$ (no serial correlation). To see this, we take the mathematical expectation for equation (1) and, accordingly, u_i can be written as

$$u_i = Y_i - E[\mathbf{Y}_i \mid \mathbf{X}1_i, \mathbf{X}2_i, ..., \mathbf{X}k_i] \qquad (10)$$

and so the distribution of u_i looks as

u_i	Possibilities of Values of u_i
$1 - E[Y_i \mid X1_i, X2_i, ..., Xk_i]$	π_1
$2 - E[Y_i \mid X1_i, X2_i, ..., Xk_i]$	π_2
...	...
$c - E[Y_i \mid X1_i, X2_i, ..., Xk_i]$	π_c

with π_1, π_2, ..., and π_c are the possibilities of obtaining the corresponding values of u_i (same as the probability of obtaining, respectively, the values of 1, 2, ..., and c for Y_i. Therefore, the variance of u_i becomes

$$Var(u_i) = E[u_i - (E(u_i))^2] = (E(u_i))^2$$

$$= \pi_1(1-M)^2 + \pi_2(2-M)^2 + ... + \pi_c(c-M)^2 \qquad (11)$$

where $E(u_i)=0$, by assumption, and M is $E[Y_i \mid X1_i, X2_i, ..., Xk_i]$. This variance depends on the conditional expectation of Y, which, of course, depends on the values taken by X1, X2,..., and Xk. Thus, ultimately the variance of u_i is heteroscedastic.

Again, the problem of heteroscedasticity is not insuperable; and even with its presence the OLS estimators are still unbiased, though not efficient (Johnston and DiNardo (2001)). One way of resolving the heteroscedasticity problem is to transform all the variables by dividing both sides of models (1) by

$$\sqrt{\mathrm{Var}(u_i)} = \text{say}, \sqrt{w_i} \qquad (12)$$

and according to Johnston(1984), Draper and Smith(1998), Maddala and Lahiri(2009) and others, we can proceed in two steps. The first one is by running the OLS method, despite the heteroscedasticity, to obtain the fitted values, which are the estimates for $E[Y_i \mid X1_i, X2_i, ..., Xk_i]$, and hence w_i. The second step is to use the estimated w_i to transform the variables and run
the OLS method.

Before we apply these two steps, we need to mention that the LPM still faces a logical problem. This is where there is no guarantee that the fitted values of Y_i will lie within the limits (1 to c in our case), despite a priori that the conditional possibility $E[Y_i \mid X1_i, X2_i, ..., Xk_i]$ must fulfil this restriction. What we could do to overcome this obstacle is to estimate the LPM by the usual OLS method and to find out whether the fitted values lie between the bounds (of 1 and c). If some are less than 1, the fitted values are considered to be 1 for those cases; and if they are greater than c, they are assumed to be equal c.

4. Cross Model Validation

Cross-validation is a method of evaluating given models by means of their predictions and to choose a model with the minimal error. We use here two forms of model validations: the `data splitting' form and the `leave one out' form. In the data splitting form, the whole data set is to be randomly split into two subsets[3]: the `estimation sample' and the `test sample'. The LPM is to be carried out on the estimation sample and then applied to the test sample to forecast the values of the dependent variable **Y** there. In the leave one out form the **Y** value for each case is set aside and the LPM is to be estimated on the remaining (n-1) data points. The prediction is then made for the case which was left out. Thus n prediction equations are derived and n **Y** values are predicted.

If y_i is a prediction of y_i, then we let $L(y_i,y_i)$ to be the loss function (Hjorth(1994)). Accordingly, we define the cross-validation error rate (CV_{ER}) for the splitting data form as

$$CV_{ER} = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \hat{y}_i) \qquad (13)$$

where m is the size of the test sample and

$$L(y_i, \hat{y}_i) = \begin{cases} 0 \text{ if } y_i = \hat{y}_i \\ 1 \text{ if } y_i \neq \hat{y}_i \end{cases} \qquad (14)$$

with the predicted values of y_i to be rounded to the nearest integer value. For

the leave one out form, the cross-validation error rate (CV_{ER}) equals

[3]Although we let this split be even in this study, however, it does not necessarily be so.

$$CV'_{ER} = \frac{1}{n}\sum_{i1}^{n} L(y_i, \hat{y}_{-i}) \qquad (15)$$

where y_{-i} is the (rounded) predicted value for subject i when it was not used in estimating the LPM and

$$L(y_i, \hat{y}_{-i}) = \begin{cases} 0 \text{ if } y_i = \hat{y}_{-i} \\ 1 \text{ if } y_i \neq \hat{y}_{-i} \end{cases} \qquad (16)$$

If we square the difference between y_i and y_i (and between y_i and y_{-i}), alternatively, we get what Maddala and Lahiri(2009) and others called the PRESS (predicted sum of squares) which is given by

$$PRESS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \qquad (17)$$

for the data splitting form and by

$$PRES\,S' = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2 \qquad (18)$$

for the leave one out form. Dividing the PRESS value by the corresponding sample size gives the cross-validation index (CV_I), and so

$$CV_I = \frac{PRESS}{m} \qquad (19)$$

and

$$CV'_I = \frac{P'RESS}{n} \qquad (20)$$

for the data splitting form and the leave one out form, respectively.

The preferred model, therefore, is the one with the smallest cross-validation error rate and/or cross-validation index.

5- Fitting The LPM For the Academic performance of Students

We consider the data of our study which are collected from a randomly selected sample of 182 students from the Omdurman Islamic University. The data are cross-classified according to the academic performance of the students, which is considered as a response, and three other categorical variables. The classification of the academic performance

is based on the cumulative rate out of 5, that is: a rate more than or equals 4.00 is considered "superb", a rate more than or equals 3.00 is considered "good", a rate more than or equals 2.00 is considered "fair", a rate more than or equals 1.00 is considered "weak", and a rate less than 1.00 is considered " very weak". The categories are coded as 1, 2, 3,4, and 5 for the 'very weak', 'weak', 'fair', 'good', and 'superb', respectively. The other three categorical variables are: the specialization of the students (social or natural sciences), whether the students live with their families or not, and the educational level of their guardians (primary or lower, intermediate, secondary, and university or higher level). It is clear that the categories of the academic performance are ordered and so the categories of the educational level of the guardian are ordered too. So, according to model(1), we have

$X1^1$ represents the situation where the specialization of the student is social science

$X2^1$ represents the situation where the student lives with his family

$X3^1$ represents the situation where the educational level of the student's guardian is university or higher

$X3^2$ represents the situation where the educational level of the student's guardian is secondary

$X3^3$ represents the situation where the educational level of the student's guardian is intermediate

In applying the OLS method to the LPM in (1), we obtain the results in table(2). The overall statistical significance of the model is indicated by the F-value of 2.25 which has a p-value of 0.051. And according to the partitioning of the regression sums of squares of the analysis of variance, most of the contribution came from $X3^1$ (56.6%) and $X2^1$ (42.8%). The coefficient of determination, R^2, indicates that only 6.0% of the variations in Y is explained by the variations in the regressor variables all together. The intercept of 3.782 indicates that, ignoring all the independent variables, the estimated value for Y would be 3.782 (which is roughly `good'). Along with the intercept coefficient, the parameters related to X11 and X32 are statistically significant at the 5% level (according to the attached t-values and their corresponding p-values). On the other hand, parameters related to X21, X31 and X33 turned out to be nonsignificant.

Table(2): The OLS Estimated LPM Of Model(1) For Y On X11, X21, X31, X32, And X33; Minitab Output.

```
The regression equation is
Y = 3.782 - 0.363 X11 - 0.132 X21 + 0.217 X31
          + 0.457 X31 + 0.018 X33
Predictor      Coef       Stdev    t-ratio         p
Constant     3.7816      0.2073      18.24     0.000
X11         -0.3631      0.1495      -2.43     0.016
X21         -0.1320      0.1541      -0.86     0.393
X31          0.2172      0.2412       0.90     0.369
X32          0.4565      0.2285       2.00     0.047
X33          0.0180      0.2302       0.08     0.938

s = 0.9869      R-sq = 6.0%      R-sq(adj) = 3.3%

Analysis of Variance
SOURCE       DF          SS         MS       F      p
Regression    5     10.9619     2.1924    2.25  0.051
Error       176    171.4118     0.9739
Total       181    182.3736

SOURCE       DF       SEQ SS
X11           1       4.6951
X21           1       0.0568
X31           1       0.0000
X32           1       6.2040
X33           1       0.0060
```

Ignoring all other variables, the differential effect of X11, which is (-0.363), shows that a unit increase in X11 results in a decrease of 0.363 in the value of Y. This means that the academic performance of students tends to be less by 0.363, on average, for the students whose specialization is social science than those specialization is natural science. On the other hand, the coefficient of 0.457 attached to the variable X32 means, holding all other variables constant, the academic performance of students is higher by 0.457, on average, for the students whose the educational level of their guardians is secondary compared with whose the educational level of their guardians is primary or lower.

The expected possibility of Y for the social science specialization of students, for those whose who live with their family, and for those whose the educational level of their guardian is university or higher is 3.504 ($\beta_0+\beta_{11}+\beta_{21}+\beta_{31}$). On the other hand, the expected possibility of Y for the natural science specialization of students, for those whose who do not live with their families, and for those whose the educational level of their guardian is secondary appears to be as 4.239 ($\beta_0+\beta_{32}$). The complete expected possibility set of Y for the different categories of the three categorical regressor variables are given in table(3).

Table(3): The OLS Estimated LPM Model Of (1) & The Expected Possibility Values of Y for the Different Categories of the Three Categorical Regressor Variables.

Educational Level of the Guardian	Living with the Family			
	Yes		No	
	Specialization		Specialization	
	Social	Natural	Social	Natural
University +	3.504	3.867	3.636	3.999
Secondary	3.744	4.107	3.876	4.239
Intermediate	3.305	3.668	3.437	3.800
Primary -	3.287	3.650	3.419	3.782

The basic question, however, before accepting these estimates and the above model is: Can we trust the estimated standard errors (and hence the t-values) reported in table(2)? The answer depends generally on the existence or otherwise of heteroscedasticity. The estimated variances of u_i, given by formula (11), are all found concentrated around 1 (the largest is 1.1149 and the smallest is 1.0011). It makes no difference dividing the variables by the square root of this variance and hence there is no need to bother about or to correct for the heteroscedasticity. To see this, table(4) gives the weighted least squares (WLS) using the two-stage procedure outlined earlier.

Table(4): The Weighted Least Squares For Model (1) For Y On X11, X21, X31, X32, And X33; Minitab Output.

```
 The regression equation is
Y' = 3.792'- 0.367 X11' - 0.143 X21' + 0.214 X31'
         + 0.470 X32' + 0.024 X33'

Predictor       Coef       Stdev    t-ratio         p
constant'     3.7924      0.2073      18.29     0.000
X11'         -0.3668      0.1511      -2.43     0.016
X21'         -0.1428      0.1551      -0.92     0.359
X31'          0.2143      0.2415       0.89     0.376
X32'          0.4701      0.2315       2.03     0.044
X33'          0.0242      0.2308       0.11     0.916

s = 0.9607

Analysis of Variance
SOURCE       DF          SS        MS        F       p
Regression    6     2436.49    406.08   439.97   0.000
Error       176      162.44      0.92
Total       182     2598.93

' denotes that the variable is divided by the weight √w
```

The conventional R^2 is not reported in table(4) and this is mainly because the intercept term is not included in the model. For this model, the sums of squares are not adjusted from the means and so, the F-value seems to be inflated accordingly.
However, the parameter estimates and their standard errors (and accordingly the t and p-values) along with the standard errors of the estimates thus obtained do not differ substantially from those obtained without the correction for heteroscedasticity. Accordingly, we retain the OLS estimates given in table(2).

The model in table(2), however, is still not convincing. The coefficients related to X21, X31, and X33 are not statistically significant, and so these variables are far from being important in the model. Including unimportant variables increases the standard errors of all estimates without improving prediction. We use the variables selection procedure to build a concise model that includes, potentially, the important variables.

The procedures of stepwise, forward, and backward elimination all reached the same conclusion and selected the model that includes X11 and X32 as a recommended model. The model is shown in table(5). It has an ample observed significance for the overall F-test (with a p-value of 0.008). The standard error for the coefficient related to X32 is now much lower compared with that in table(2) and, as a consequence, the corresponding t-values are now higher.

Table (5): The Final (OLS) Output Using the Stepwise, Forward, and Backward

Multiple R	.22781				
R Square	.05190				
Adjusted R Square	.04130				
Standard Error	.98284				
Analysis of Variance					
	DF	Sum of Squares	Mean Square		
Regression	2	9.46461	4.73230		
Residual	179	172.90902	.96597		
F = 4.89901	Signif F = .0085				
Variables in the Equation					
Variable	B	SE B	Beta	T	Sig T
X32	.348412	.156797	.162181	2.222	.0275
X11	-.347889	.147014	-.172714	-2.366	.0190
(Constant)	3.791050	.107661		35.213	.0000

Procedures for Model (1) Of Y On X11, X21, X31, X32, and X33; SPSS Output.
The estimated R^2 of 5.2% is seen pretty low and this might look to contradict the F-value (which in turn tests the significance of the R^2). The fact is, however, the R^2 is unlikely to be high in the case of the LPM, since the response variable takes only limited values and the scatter plot of this variable with any of the independent variables is expected to be concentrated on those limited values and this results in low partial correlations and, accordingly, low multiple correlation. It is not surprising, therefore, to see the R^2 in our case as much low as 5.2% and to be significant at the same time.

Based on this refined model, the conditional possibility of Y is lower by 0.3484 (a third) for all the students whose specialization is social science. Likewise, the conditional possibility of Y is higher by 0.3479 (a third) for all the students whose the educational level of their guardian is secondary. For those whose the educational level of their guardian is secondary and their specialization is social science, the increment in the conditional possibility of Y is almost 0. For those whose their specialization is natural science (i.e., when ignoring the coefficient of X11) and the educational level of their guardian is primary or lower, the conditional possibility of Y is 3.791, which is "good".

6- The Diagnostic Checks

Compared with the all variables models (OLS & WLS) the stepwise, forward, and backward procedures shown in table(5) is obviously the best. However, for a similar model reached by the stepwise, forward, and backward procedures (with the same selected predictor variables) we need to cross validate the models to assess how well they predict in an independent sample(s) of data. In other words, to determine which of the two models have more generalizability. table(6) along with table(7) the cross-validation rates and indices for these two models across the three samples: the data split sample, the leave one out sample, and the whole data set sample.

The cross-validation error rates in table(6) seem to be considerably higher than what we usually expect for all the models in the three samples. About two-thirds of the Y scores are incorrectly predicted with the OlS model and slightly lower than that with the WLS model. However, for a dependent variable like Y which has several outcomes (superb, good, fair, weak, and very weak), it is more likely that a predicted value will result in a mismatch than when this variable has only two outcomes (yes and no, say). This is probably because the chance for the outcome to be correctly forecasted will be smaller as the number of outcomes increase, that is, 4 out of 16 for our case compared with a 2 out of 4 chance if Y has a binary outcome.

Table(6): The Cross-validation Error Rates for the Stepwise, Forward, and Backward Procedures OLS Model Of Table(5) and WLS Model.

Validation Sample	The OLS Model	The OLS Model
Data Split	68.1%	63.7%
Leave One Out	65.4%	64.4%
Whole Sample	65.4%	67.4%

Table(7) considers the error rates with a minor mismatch being allowed, for instance, the `good' outcome to be predicted as `fair' but not as `weak' or `very weak', i.e., one-level mismatch. For the two models, the error rates are now remarkably low compared with the previous ones in table(6), especially with the whole sample and the leave one out sample. It is also noted that OLS model has now relatively lower error rates than its counterpart, the WLS model.

Table(7): The Cross-validation Error Rates for the Stepwise, Forward, and Backward Procedures OLS Model Of Table(5) and WLS Model, Allowing for Minor Mismatch.

Validation Sample	The OLS Model	The OLS Model
Data Split	13.2%	14.5%
Leave One Out	11.5%	11.7%
Whole Sample	11.5%	11.1%

Table(7) considers the error rates with a minor mismatch being allowed, for instance, the `good' outcome to be predicted as `fair' but not as `weak' or `very weak', i.e., one-level mismatch. For the two models, the error rates are now remarkably low compared with the previous ones in table(6), especially with the whole sample and the leave one out sample. It is also noted that OLS model has now relatively lower error rates than its counterpart, the WLS model.

In table (6) we notice that no major difference in the error rates between the split sample and whole data sample (in fact the split sample has smaller error rates than the whole sample for the WLS). This is possibly because both of them are large. For the leave one out sample and the

whole sample, the error rates in tables(6) and (7) are quite identical for the OLS and we would expect this, since the difference in the sample size is one. This is, therefore, a further possible evidence to keep hold of the OLS model of table (5).

7- Summary

In this research we used the LPM techniques to analyze categorical data with ordered categories for the response variable. The technique looks simple and can be applied by the familiar OLS or WLS methods. By simple we mean the conditional mean value of the response variable is simply the conditional possibility of the event, given the values of the other categorical regressor variables. However, although simple to apply as we said, this model has three main problems: nonnormality and heteroscedasticity of the error term as well as the possibility of fitted values lying outside the categories of the response variable. The research showed how these three problems can be resolved to allow the model to give straightforward interpretations similar to that given by the usual regression analysis.

For an application of the technique, we used a random sample data of 182 students from the Omdurman Islamic University. The data are cross-classified according to the academic performance of the students, which is considered as a response for three other categorical variables. The classification of the academic performance are ‘superb’, ‘good’, ‘fair’, ‘weak’, and ‘very weak’. The other three categorical variables are: the specialization of the students (social or natural sciences), whether the students live with their families or not, and the educational level of their guardian (primary or lower, intermediate, secondary, and university or higher level).

The result showed that the conditional possibility of the academic performance of the students is lower by 0.3484 (a third) for all the students whose specialization is social science. Likewise, the conditional possibility of the academic performance of the students is higher by 0.3479 (a third) for all the students whose the educational level of their guardian is secondary. For those whose the educational level of their guardian is secondary and their specialization is social science, the increment in the conditional possibility is almost 0. For the students whose their specialization is natural science and the educational level of their guardian is primary or lower the conditional possibility of their academic performance is 3.791, which is "good". The variable related to the situation where the students live with their families or not seemed not to be statistically significant in affecting their academic performances.

8- References

1. Adam, Amin. I. (1993). Analysing Categorical data: Various Solutions and Different Conclusions. Young Statistical Meeting, Liverpool.
2. Adam, Amin. I. (1996). Analysis of Categorical Data from a Case Study of Child Safety. Unpublished PhD thesis: University of Keele, Dept. of Mathematics, England.
3. Adam, Amin. I. (2010). Concepts on the Chi-square Test of Independence for Analyzing Categorical data. Journal of the Faulty of Economics & Political Science, Omdurman I. University, Sudan, Vol.4:131-143.
4. Adam, Amin. I. (2010). Local-Local, Local-Global, Global-Local and Global-Global Odds Ratios for Categorical data. J. of Economics and Political and Statistical Sciences, Omdurman I. University, Vol. 5:165-186.
5. Adam, Amin. I. (2010). Measures of Associations for Ordered Categorical Data: Different Measures but Similar Conclusions. J. of Economics and Political and Statistical Sciences, Omdurman I. University, Vol. 6:180-198, December 2010.
6. Agresti, A. (2002). Categorical Data Analysis, 2nd ed. Wiley, New York.
7. Agresti, A. (2007). An Introduction to Categorical Data Analysis. Wiley, New York.
8. Agresti, A. (2010). Analysis of Ordinal Categorical Data. Wiley, New York.
9. Baglivo, J., Oliver, D. & Pagano, M. (1992). Methods for Exact Goodness-of-Fit Tests. Journal of the American Statistical Association 87:464-469.
10. Becker, M. P. & Clogg, C. C. (1989). Analysis of Sets of Two-Way Contingency Tables Using Association Models. Journal of the American Statistical Association 84:142-151.
11. Bilder, C. & Loughin, T. M. (2007). Modeling Association Between Two or More Categorical Variables that Allow for Multiple Categorical Choices. Communications in Statistics 36:433-451.
12. Bower, K.M. (2000), "Analysis of Variance (ANOVA) Using MINITAB" Scientific Computing & Instrumentation.
13. Draper, N. R. & Smith, H. (1998). Applied Regression Analysis, 3rd ed. Wiley, New York.
14. Everitt, B. S. (1977). The Analysis of Contingency Tables. Chapman & Hall, London.
15. Eye, A. V. & Bogat, G. A. (2009). Analysis of Intensive Categorical Longitudinal Data. Springer, New York.
16. Fan, Y. (2008). Strategic Groups and cluster Analysis. Henry Stewart, London.
17. Fienberg, S. E. (2007). The Analysis of Cross-classified Categorical Data. Springer, New York.
18. Freeman, D.H. (1987). Applied Categorical Data Analysis. Marcel Dekker, New York.
19. Greenland, S. (1991). On the Logical Justification of Conditional Tests for Two-by-Two Contingency Tables. American Statistician 45:248-251.
20. Gujarati, D. N. (2004). Basic Econometrics, 4th ed. McGraw-Hill, New York.
21. Hjorth, J. S. U. (1994). Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap. Chapman, London.
22. Imrey, P. B. & Koch, G. G. (2005). Categorical Data Analysis. Wiley, New York.
23. Johnston, J. (1984). Econometric Methods, 3rd ed. McGraw-Hill, New York.
24. Johnston, j. & DiNardo, J. (2001). Econometric Methods, 4th ed. McGraw-Hill, New York.
25. Liu, I. & Agresti, A. (2005). The Analysis of Ordinal Categorical Data: An Overview and a Survey of Recent Development. Sociedad de Estadistica e Investigacion OperativaTe Vol.14 No. 1:1-73.
26. Maddala, G. S. & Lahiri, K.(2009). Introduction to Econometrics. Wiley, New York.
27. McCullagh, P. (1980). Regression Models for Ordinal Data. J. Roy. Statist. Soc. B 42:109-142.

28. Ott, R. L. & Longnecker M. (2008). An Introduction to Statistical Methods and Data Analysis, 6th ed. Brooks/Cole, Bolmont, U.S.A.
29. Powers, D. A. (2008). Statistical Methods for Categorical Data Analysis, 2nd ed. Emerald, Bingley, U.K.
30. Simono, J.(2003). Analyzing Categorical Data. Springer, New York.