ISSN No:-2456-2165

Analyzing Likert Scale Data using Cumulative Logit Response Functions with Proportional Odds

Isaac Oluwaseyi Ajao and Aladesuyi Alademomi
Department of Mathematics and Statistics,
The Federal Polytechnic, Ado-Ekiti, Ado-Ekiti, Nigeria

Abstract:- Handling simple binary response data with logistic regression has solved many problems encountered in data analysis across various walks of life. However, dealing with ordinal responses, especially when they are more than two levels has remained a big challenge to researchers. This paper therefore focuses attention on the application of cumulative logit response function with proportional odds in order to show its robustness over chi-square, t-tests, percentages and so on, in analyzing likert scale data which is common among users of statistics in social sciences, environmental, and medical sciences. To implement this, 500 random observations on five socio-demographic variables were simulated.In order to justify the use of proportional odds, score test was carried out on the data, and the assumption was not rejected at 5% level (p-value = 0.4222), this justifies the use of the method. Also, the Deviance and Pearson goodness of fit statisticsshow p-value = 0.7326 and 0.8130 respectively, this reveals that the model fits the data adequately. Moreover, the proportional odds model is fitted and reliable predictions are made. The method is robust for analyzing ordinal response data such as likert scale.

Keywords:- *Ordinal data, likert scale, proportional odds model, logistic regression.*

I. INTRODUCTION

Ordinal logistic models are utilized to obtain relationship between an ordered response variable and a set of predictors. An ordinal variable is a variable that is in categories and ordered, for example, "poor", "great", and "superb", which may demonstrate an individual's present wellbeing status or the maintenance record of a vehicle. Assessments from studies can be requested likewise, "unequivocally concur", "concur", "dissent" and "emphatically conflict". This entry is concerned distinctly with multiple results. This entry is concerned distinctly with models in which the results can be ordered.

In ordinal logit, a hidden score is calculated as a linear function of the predictors and a set of benchmarks. The likelihood of noticing result *i* compares to the likelihood that the assessed straight capacity, in addition to irregular blunder, is inside the scope of the benchmarks assessed for the result$Pr(outcome_j = i) = Pr(\kappa_{i-1} < \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + u_j \leq \kappa_i)$

$u_j$is thought to be strategically disseminated in arranged logit. Regardless, we calculated the coefficientsβ_1, β_2, . . .,$\beta_k$along with the benchmarksκ_1, κ_2, . . ., κ_{k-1}, where *k* is the number of possibleoutcomes. κ_0 is taken as -∞, and $\kappa_k$is taken as +∞

Long and Freese (2014) examined models for ordinal results and their application utilizing STATA Cameron and Trivedi (2005) represent multinomial models, including the model fit. At the point when you have a subjective ward variable, a few assessment methodology are accessible. A famous decision is multinomial strategic relapse, yet in the event that you utilize this technique when the reaction variable is ordinal, you are disposing of data since multinomial logit disregards the arranged part of the result. Requested logit and probit models give a way to misuse the requesting data. There is mutiple "requested logit" model. The model fit which we will call the arranged logit model, is otherwise called the relative chances model. All arranged logit models have been determined by beginning with a parallel logit/probit model and summing it up to take into account in excess of two outcomes.The relative chances requested logit model is supposed on the grounds that, on the off chance that we think about the chances,odds(k) =P ($Y \leq k$)/P ($Y > k$), then odds(k_1) and odds(k_2) have similar proportion for all free factor mixes. The model depends on the rule that the lone impact of consolidating abutting classifications in arranged absolute relapse issues ought to be a deficiency of proficiency in assessing the relapse boundaries (McCullagh 1980). This model was likewise portrayed by McKelvey and Zavoina (1975) and, already, by Aitchison and Silvey (1957) in an alternate mathematical structure. Brant (1990) offers a bunch of diagnostics for the model. Peterson and Harrell (1990) recommend a model that permits nonproportional chances for a subset of the informative factors. The generalization model oddballs the guideline on which the arranged logit model is based. Anderson (1984) contends that there are two unmistakable sorts of requested straight out factors: "assembled constant", like pay, where the "type a" model applies; and "evaluated", like degree of help with discomfort, where the generalization model applies. Greenland (1985) freely built up a similar model. The generalization model beginnings with a multinomial strategic relapse model and forces imperatives on this model. Integrity of fit for ordinal calculated relapse can be assessed by contrasting the probability esteem and that got by fitting the multinomial strategic relapse model. Leave $\ln L_1$ alone the log probability esteem revealed by ordinal strategic relapse model, and let $\ln L_0$ be the log-probability esteem detailed by multinomial calculated relapse model. On the off chance that there are p free factors (barring the consistent) and k classes, multinomial strategic relapse model will gauge p(k-1) extra boundaries. We would then be able to play out a "probability proportion test" that is, calculate -2($\ln L_1$- $\ln L_0$),

ISSN No:-2456-2165

and compare it with $\chi^2\{p(k - 2)\}$.This test is intriguing simply because the ordinal logistic model isn't settled inside the multinomial logistic model. An enormous worth of-2($\ln L_1$- $\ln L_0$) should, however, be taken as evidence of poorness of fit.Marginally large values, on the other hand, should not be taken too seriously.The coefficients and cut-points are estimated using maximum likelihood.

In our parameterization, no constant appears, because the effect is absorbed into the cut-points. ordered logistic model and ordered probit model begin by tabulating the response variable. Category i= 1 is defined asthe minimum value of the variable, i= 2 as the next ordered value, and so on, for the empiricallydetermined k categories.

II. MATERIALS AND METHODS

Binary logistic regression is a summed up linear model that utilizes the binomial circulation and a logit connect work. At the point when your reaction has multiple levels, you utilize a multinomial dispersion and diverse connection capacities to address the idea of the reactions, and distinctive linear indicators to demonstrate the probabilities. For the accompanying conversation, assume our reaction variable Y has J = 4 levels (1, 2, 3, and 4) for instance 4-point likert scale, so we write $\pi_{ij} = \Pr(Y_i = \text{j})$

III. LINK FUNCTIONS

A. Generalized Logit Link

Characterize a link function so every response function stands out a lower level from the last level: $log\left(\frac{\pi_{i1}}{\pi_{i4}}\right)$ $log\left(\frac{\pi_{i2}}{\pi_{i4}}\right)$ $log\left(\frac{\pi_{i3}}{\pi_{i4}}\right)$

This is the summed up logit link, and it disregards the request for the responses, past recognizing the final remaining one as the reference response. You utilize this link when you have ostensible information.

B. Cumulative Logit Link

The most famous ordinal connection work utilizes each likelihood in each capacity by differentiating the lower levels of Y with the more elevated levels of Y. Leave the aggregate likelihood alone meant as

$\theta_{ij} = \Pr(Y_i \leq \text{j})$:

$\text{logit}(\theta_{i1}) = log\left(\frac{\pi_{i1}}{\pi_{i2}+\pi_{i3}+\pi_{i4}}\right)$, $\text{logit}(\theta_{i2}) = log\left(\frac{\pi_{i1}+\pi_{i2}}{\pi_{i3}+\pi_{i4}}\right)$, $\text{logit}(\theta_{i3}) = log\left(\frac{\pi_{i1}+\pi_{i2}+\pi_{i3}}{\pi_{i4}}\right)$

This is the aggregate logit link. As we move from the first logit function to the second and from the second logit function to the third, the numerator increments and the denominator diminishes, so the aggregate logits are expanding.

C. Proportional Odds Model

The overall model has inconsistent inclines for the indicators, and we really want sufficient information to appraise an alternate coefficient for every indicator in every response function. To work on this model, we can compel a requesting on the straight indicators by utilizing a similar slant boundaries for every response function and by obliging the captures to increment $(\alpha_1 < \alpha_2 < \alpha_3)$or decline:

$$g(j) = \alpha_j + \text{X}'\beta \qquad j = 1, 2, 3$$

This model has $J-1+p$ boundaries and accordingly requires less information for a sufficient fit than the overall model requires. It additionally gives a more direct understanding. Figure the distinction of the jth response function between two subpopulations h and I to see the effect of this model:

$$g(hj) - g(ij) = (x_h - x_i)'\beta \qquad j = 1, 2, 3$$

This distinction is relative to the distance between the logical factors, and the thing that matters is a similar regardless of which response function you consider. This is the equivalent slants supposition, which is additionally called the equal lines suspicion. One can apply the equal lines presumption to any of the link functions, yet it is most ordinarily utilized with the aggregate logit link. At the point when we utilize the aggregate logit link, the supposition that is the corresponding chances suspicion, the model is the relative chances model, and the distinction of combined logits (g) is the log total chances proportion. Of course, when playing out the examination utilizing SAS, the relative chances model is fitted and it is joined with the combined logit link when we have multiple response levels.

IV. DATA AND DATA ANALYSIS

The data used was obtained through simulation. All data simulation and analyses were done using *Excel* and *SAS studio 3.5*. The data is on factors influencing decisions made by final year National Diploma students on whether to go further their education in the university or go back for HND programme. The response variable *decision* has 3 levels which includes *very sure*, *sure*, and *not sure*. The predictors are parental education (*paredu*), friends' education (*fredu*), opinion on university degree (*degree*), and grade point average (*gpa*), each having 2 levels.

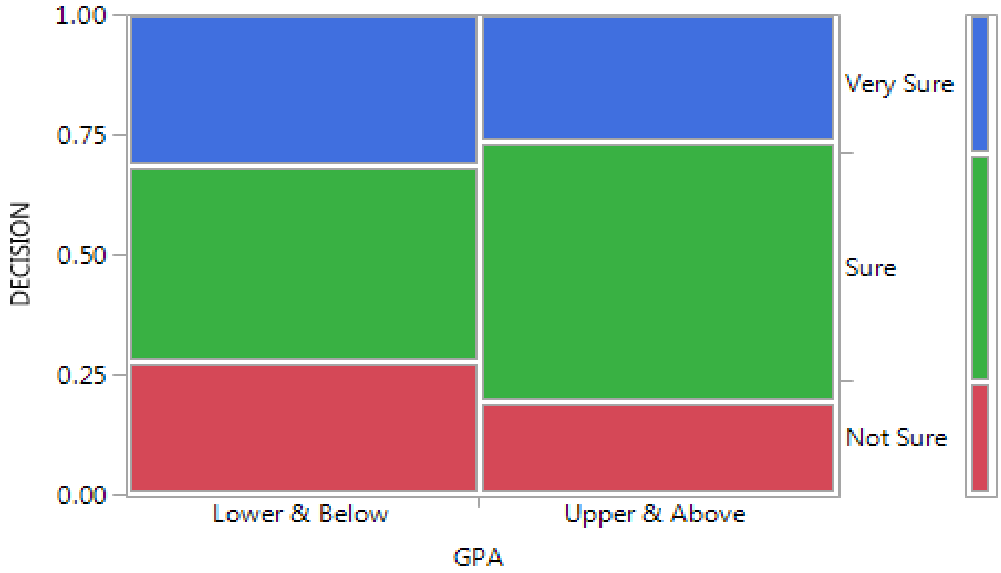

Fig. 1: Mosaic plot of Decision by GPA

The proportions on the x-axis represent the number of observations for each level of the X variable, which is GPA. The proportions on the y-axis at right represent the overallproportions of Not sure, Sure, and Very sure for the combinedlevels (Lower & Below and Upper & Above).Very few Upper & Above students fall into the Not sure category, majority of them are sure.The majority of theLower & Belowstudents fall into the sure categories, but not as high as the students in Upper & Above. The proportion of Lower & Above students who are not is higher than those in Upper & Above.The proportion of students who are sure of their decision is higher.

Table 1: Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
4.8297	4	0.3052

Table 1 displays a score test for the proportional odds assumption; the test does not reject the null hypothesis that the proportional odds assumption holds. This score test actually tends to reject the null hypothesis more often than it should; Stokes,Davis, and Koch (2012) say that this statistic needsapproximately five observations (or frequencies) for eachoutcome at each level of each main effect, because small samples might make the statistic artificially large.This score test is a good confirmatory test if it does not reject the null; however, if it rejects the null, then we need other means to justify the proportional odds assumption.

Table 2: Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	23.3424	26	0.8978	0.6135
Pearson	23.1745	26	0.8913	0.6231

It is confirmed in table 2 that the model fits the data adequately, since the hypothesis that the fit is good is not rejected

Table 3: Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1056.619	1056.714
SC	1065.049	1082.002
-2 Log L	1052.619	1044.714

The fit statistics that are shown in table 3 are often used to compare nested models. The difference of the–2 Log L statistics forms the likelihood ratio statistic that is shown in table 4

Table 4: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.9051	4	0.0951
Score	7.7908	4	0.0996
Wald	7.9109	4	0.0949

ISSN No:-2456-2165

The three global tests that are displayed in table 4 evaluate the significance of all the predictors combined.They tell us only whether the model has some significance; they don't say anything about the effect ofindividual predictors. Tests for parameters being jointly zero are not rejected. This means that the predictors areunrelated to students' decision making; that is, the model cannotexplain significant amount of variation in the data.

Table 5: Type 3 Analysis of Effects

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
PAREDU	1	0.3221	0.5703
FREDU	1	0.3585	0.5493
DEGREE	1	0.8866	0.3464
GPA	1	6.4709	0.0110

Tests that the parameters for a class effect are all zero are not rejected except for GPA

Table 6: Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-0.6255	0.1926	10.5521	0.0012
Intercept	1	1	0.3888	0.1911	4.1400	0.0419
PAREDU	1	1	-0.0963	0.1697	0.3221	0.5703
FREDU	1	1	0.1011	0.1689	0.3585	0.5493
DEGREE	1	1	-0.1589	0.1688	0.8866	0.3464
GPA	1	1	-0.4294	0.1688	6.4709	0.0110

The fitted model is:

Table 7: Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
PAREDU 1 vs 0	0.908	0.651	1.267
FREDU 1 vs 0	1.106	0.795	1.541
DEGREE 1 vs 0	0.853	0.613	1.188
GPA 1 vs 0	0.651	0.468	0.906

The interpretation of these odds ratios is that the odds of a student very sure of going to the university after ND among parents who have degrees is 0.908 times higher than among whose parents are not graduates. The odds of students who are very sure of proceeding to the university among friends who are students in the university is 1.106 times higher thanof those who do not have. Similarly, those who are of opinion that schools have the same degree have 0.853 times higher than who don't believe in it. Lastly, the odds of students who are sure of entering the university among having upper & higher grade is 0.651 times higher than among those having lower grades.

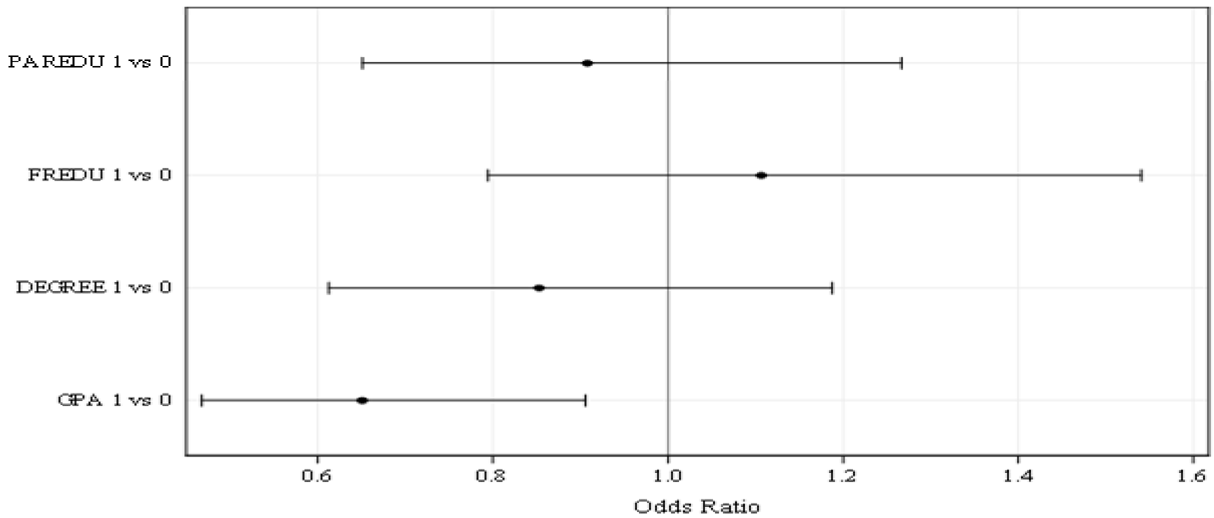

Fig. 2: Odds Ratios with 95% Wald Confidence Limits

The odds ratio estimates presented in table 7 are represented in fig. 2. It can be seen that from the chart that the odds ratio is highest in FREDU and least in GPA.

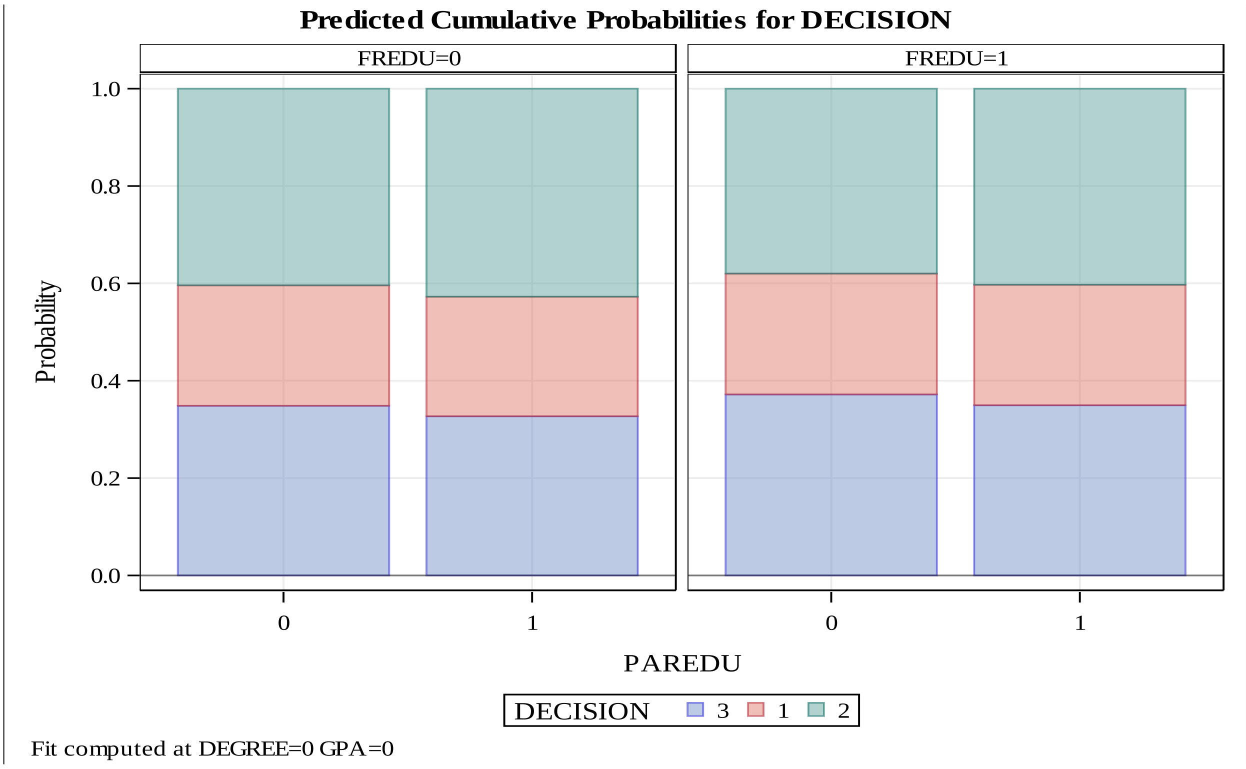

Fig. 3: Predicted cumulative probabilities for DECISION from FREDU and PAREDU

The predicted probability of a student who is“very sure” is the lowesthaving friends who are not in the universityand those who are there with parents who not university graduates.A studentwho is“very sure” has the lowest predicted probabilities in both cases, but the impact of parent having university degree has positive influence by the observed increase in the predicted probability of “sure” category. On the other hand, since the student’s friend is having a university degree the predicted probabilities has increased.

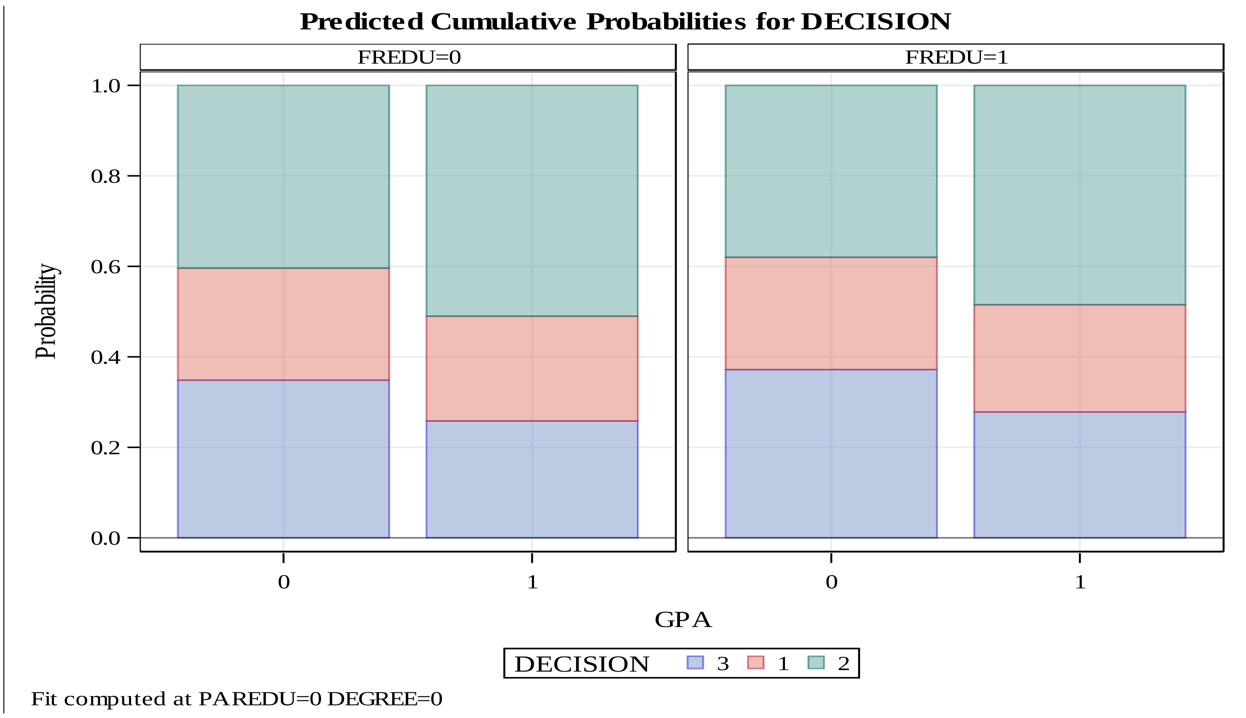

Fig. 4: Predicted cumulative probabilities for DECISION from FREDU and GPA

The effect of the students’ GPA is significant in the fitted probabilities as also revealed in table 5. The probabilities are of students who are “sure” attending the university is generally high especially among students those who have “upper &above” grade.

Table 8: Results using the chi-square

variables	chi-square	p-value
decision &paredu	0.4096	0.8150
decision &gpa	9.2391	0.0100
decision &fredu	1.2624	0.5320
decision & degree	1.4170	0.4920

According to table 8, the chi-square test only reports the significance or non-significance of a factor. Once a factor is not significant, it means it does not have effect on the response variable, and if it is significant, one cannot measure the degree of the effect unlike the proportional odds model used above.

V. SOME IDENTIFIED WEAKNESSES OF THE CHI-SQUARE AND T-TEST WHEN USED FOR ORDINAL DATA

The chi-square cannot measure ordinal responses, it only establishes relationships, another one is that no model is fitted in chi-square test. Similarly, the t-test violates the assumptions of normality and homogeneity of variance if used for ordinal data, and lastly, the t-test cannot fit a model for ordinal data and as a result lack the ability of producing the predicted probabilities.

VI. SUMMARY OF RESULTS

It can be seen from table 1 that the test does not reject the null hypothesis that the proportional odds assumption holds. It is also established in table 2 that the model fits the data adequately. The odds of a student very sure of going to the university after ND among parents who have degrees is 0.908 times higher than among whose parents are not graduates. The odds of students who are very sure of proceeding to the university among friends who are students in the university is 1.106 times higher than of those who do not have. Similarly, those who are of opinion that schools have the same degree have 0.853 times higher than who don't believe in it. Lastly, the odds of students who are sure of entering the university among having upper & higher grade is 0.651 times higher than among those having lower grades. It can be seen from the chart that the odds ratio is highest in FREDU and least in GPA. The predicted probability of a student who is "very sure" is the lowest having friends who are not in the university and those who are there with parents who not university graduates. A student who is "very sure" has the lowest predicted probabilities in both cases, but the impact of parent having university degree has positive influence by the observed increase in the predicted probability of "sure" category. On the other hand, since the student's friend is having a university degree the predicted probabilities has increased. The effect of the students' GPA is significant in the fitted probabilities as also revealed in table 5. The probabilities are of students who are "sure" attending the university is generally high especially among students those who have "upper & above" grade.

VII. CONCLUSION

The proportional odds model has been proved better than many others such as the chi-square test and the t-test when analyzing ordered category response data like the likert scale mostly used in measuring survey data. The real contribution of each of the predictors to the outcome categories are obtained using the proportional odds model, their odds ratio estimates, and their predicted probabilities after the models were obtained. Thesestrengths which are weaknesses in the aforementioned methods make the proportional odds model better whenever analysis involving ordered categories in the responses is required.

REFERENCES

[1.] Aitchison, J., and S. D. Silvey. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*44: 131–140.

[2.] Anderson, J. A. (1984). Regression and ordered categorical variables (with discussion). *Journal of the Royal StatisticalSociety, Series B* 46: 1–30.

[3.] Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*46: 1171–1178. Cameron, A. C., and P. K. Trivedi. (2005).*Microeconometrics: Methods and Applications*. New York: CambridgeUniversity Press.

[4.] Greenland, S. (1985). An application of logistic models to the analysis of ordinal responses. *Biometrical Journal* 27:189–197.

[5.] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*42: 109–142.

[6.] McKelvey, R. D., and W. Zavoina. (1975). A statistical model for the analysis of ordinal level dependent variables.*Journal of Mathematical Sociology* 4: 103–120.

[7.] Peterson, B., and F. E. Harrell, Jr. (1990). Partial proportional odds models for ordinal response variables. *AppliedStatistics* 39: 205–217.

[8.] SAS Institute Inc. (2012). "SAS Note 37944: Plots to Assess the Proportional Odds Assumption in an Ordinal Logistic Model," http://support.sas.com/kb/37944.

[9.] SAS Institute Inc. (2013)."Paper 446-2013: Ordinal Response Modeling with the logistic procedure,"http://support.sas.com

[10.] Stokes, M. E., Davis, C. S., and Koch, G. G. (2012).*Categorical Data Analysis Using SAS*, 3rd Edition, Cary, NC: SAS Institute Inc.