

Ordered Logistic Regression on the Mental Health of Undergraduate Students

Maureen T. Nwakuya^{1,*}, Omosivie Mmaduka²

¹Department of Mathematics/Statistics, University of Port Harcourt, Port Harcourt Rivers State, Nigeria

²Department of Prevention and Social Medicine, College of Health Science, University of Port Harcourt Teaching Hospital, Port Harcourt Rivers State, Nigeria

Abstract Mental health is the state of well-being of an individual, where such individual can cope with the stresses of life and still make contribution to his/her community. Research has shown that there is an association between behavioral tendencies of undergraduate students and their mental health. This we believe is due to stress of full time academic program. Bearing this in mind we focused our research in identifying these factors that affect the mental health of students and ascertain the level of relationship between them and the student's mental health. We adopted the stratified sampling technique for data collection and used self-administered questionnaire to identify the prevalence of factors known to be associated with mental health. The ordered logistic regression was carried out to obtain a proportional odds model that was used to model this relationship. The results show that when the current well-being of the students increase by a unit the odds of mental health state of the student being in an unstable state or mildly-unstable state versus stable state increases by 36.27%, given that any other variable is held constant. Also when the coping strategy increases the odds also increases by 33.16%, we can also observe that increase in the mental health assessment also increases the odds of mental health in unstable or mildly unstable state versus stable state, by 22.49%. A close look at the odds ratio and the regression coefficients makes us conclude that the most influential factors of the mental health is the current well-being of the student, followed by coping strategies and lastly the mental health assessment.

Keywords Mental health, Ordered Logistic Regression, Stratified Sampling, Proportional Odds Ratio and Categorical data

1. Introduction

Mental health is the state of well-being of an individual, where such individual can cope with the stresses of life and still make contribution to his/her community. The World Health Organization (WHO) estimates that about 20% of Nigerians suffer mental illness with mental disorders accounting for 13% of global burden of disease. Experts have tasked the government on the need to increase its investment in mental health, pass the Mental Health Bill and have a health policy on mental health as practiced internationally. The WHO endorsed mental health as a universal human right and a fundamental goal for health care systems of all countries (WHO, 2005). Unfortunately, mental health systems in many low and middle income countries in Sub-Sahara Africa face challenges in ensuring optimal mental health services, [10].

In Nigeria, the issue of mental health has attracted little attention, much less that of undergraduate students. Often undergraduate mental health victims are stigmatized, scorned and neglected by families in the belief that their problem is beyond remedy, so many of them abandon school and are left to wander about on the streets, sleeping under the bridges and causing menace to the society. The question that comes to mind is; how many Universities in Nigeria have an office which deals with mental health, psychological and emotional issues of students? The issues of mental health illness can be nipped at the board, if preventive measures and coping strategies are laid down.

The population for this research work is University of Port Harcourt and the data size for this research work was obtained using the stratified sampling scheme, where the departments served as the strata. The total sample size is 706.

2. Ordered Logistic Regression Model

Logistic regression and Ordinary least square (OLS) have the same goal, which is to model a response variable as a function of one or more explanatory variables, the major difference is that OLS requires the response variable to be

* Corresponding author:

maureen.nwakuya@uniport.edu.ng (Maureen T. Nwakuya)

Published online at <http://journal.sapub.org/ijps>

Copyright © 2019 The Author(s). Published by Scientific & Academic Publishing

This work is licensed under the Creative Commons Attribution International

License (CC BY). <http://creativecommons.org/licenses/by/4.0/>

continuous while logistic regression requires the response variable to be a dichotomous categorical variable. Often there are occasions when the response variable is polychotomous. Such response variable can be classified into two categories-multinomial and ordinal, when the response variable is classified according to their order of magnitude, the multinomial logistic regression cannot be used rather the ordinal/ordered logistic is most appropriate, [11]. Many logistic regression models have been developed for analyzing ordinal response variable. In situations where there are many response variables, a multivariate approach is considered, but for this paper only one response variable is considered with more than two explanatory variables, so the ordered logistic regression is considered. Ordered logistic regression is an extended type of logistic regression where the response categorical variable is ordered into more than two categories. There are several ordered/ordinal logistic regression models such as Proportional Odds Model (POM), Continuous with Restrictions, Stereotype Model etc. Ordinal Logistic regression was originally proposed by [12]. The name cumulative link model is adopted from [1] and it was later changed to proportional Odds model by [9]; [2].

The work focuses on how well the response variable (Mental health state) can be predicted by the answers from questions based on five factors. This research work investigated the effects of these factors on the mental health of undergraduate students using the ordered logistic regression. The response (mental health state) was ordered into three categories. The underlying model for this work is the Proportional odds model. A major assumption underlying ordinal logistic regression is that the relationship between each pair of outcome groups is the same, that is the coefficient that describe the relationship between the lowest versus all the higher categories of the outcome variables are the same as the one that describe the relationship between the next lowest category and all the higher categories etc. This is known as the parallel regression assumption or the proportional odds assumption. To ascertain that the Proportional odds model is appropriate, this assumption must hold.

3. Proportional Odds Model (POM)

The proportional odds model is a special case of the cumulative logit model. It is not based on individual probability of each class, but on the cumulative probabilities of the class considered j and the $(j-1)$ precedent classes [9]. Proportional odds model is a model that allows separated intercepts for the cumulative logit, but restricted the parameter sets of the predictors to be same across all the logits. This model basically compares the probability of a response less than or equal to a given category ($j=1, \dots, J-1$) to the probability of a response greater than this category. In addition, this model is composed of $J-1$ parallel linear equations. The model has $(J-1)$ parameters. Let the probability that the response Y_i falls in the $j=1, \dots, J$ category

be denoted by π_{ij} ie $\pi_{ij} = \Pr\{Y_i = j\}$. We then define the corresponding cumulative probability that the response falls in the j^{th} category or below as;

$$Y_{ij} = \Pr\{Y_i \leq j\} \quad (1)$$

so we have that

$$Y_{ij} = \pi_{i1} + \pi_{i2} + \dots + \pi_{ij} \quad (2)$$

Let $g(\cdot)$ denote a link function mapping probabilities to the real line. The link function is used to transform the cumulative probabilities into linear function of the independent variables.

The logit transformation of the cumulative response probabilities γ_{ij} is given by;

$$g(Y_{ij}) = \text{logit}(Y_{ij}) = \log(Y_{ij}/1 - Y_{ij}) \quad (3)$$

And this equation is assumed to be a linear function of the explanatory variables, thus; the cumulative logit model is given by;

$$\begin{aligned} g(Y_{ij}) &= \log(Y_{ij}/1 - Y_{ij}) = \log\{[Pr(Y_i \leq j)/x_i] / [1 - Pr(Y_i \leq j)/x_i]\} \\ &= \log\{(\pi_{i1} + \pi_{i2} + \dots + \pi_{ij}) / (\pi_{i1+1} + \pi_{i2+2} + \dots + \pi_{ij})\} \\ &= \theta_j - x_i'\beta \text{ for } j = 1, \dots, J-1 \end{aligned} \quad (4)$$

This is called the proportional odds model (POM), where θ_j is a constant representing the baseline value of the transformed cumulative probability for category j , x_i is the vector of explanatory variables for the i th observation and β is the corresponding set of regression parameters. The effect of the explanatory variables is either to raise or lower the odds of a response in category j . This model was derived by [12]. Some sources write the cumulative logit model with a plus on the right-hand side, but there is a good reason for the minus. It means that larger the value of $x_i'\beta$, the higher the probability of the response falling in a category at the upper end of the response scale ie the higher value of the covariates indicates a positive association with the higher value of the response category. For small values of $x_i'\beta$ the response is likely to fall in the low category and for large values $x_i'\beta$ the response is likely to fall in the high category. A key point is that the regression part $x_i'\beta$ is independent of j , so that β has the same effect for each of the $j-1$ cumulative logits. It is worthy of note that $x_i'\beta$ does not contain an intercept, since the θ_j (threshold) act as intercepts.

Definitions of Terms:

The predictor variables are defined thus:

- ◀ Current_wellbeing:- This we define to mean how the individual perceives him/herself, in terms of being cheerful, relaxed, useful, loved etc.
- ◀ Current_Concerns:- These are the current concerns of undergraduate students, such as family, relationships, etc.
- ◀ Concerns on mental health:- This we considered as issues pertaining to someone's health like suicidal tendencies, seeing a doctor on health matters, attempt to be get better etc.
- ◀ Mental health Assessment:- This we defined to be mental stability in terms of self-control, attitude to life

etc.

- ✦ Coping Strategies:- This is defined as the measures taken in other to maintain a stable mental health,

4. Methodology

The aim of this research work is to investigate the relationship between the mental health state of undergraduate students and factors that affect the mental health of these students. This research work utilized data gotten from 706 respondents through questionnaire method. We considered a response variable Y (mental health status) with three categories, namely; Stable, Mildly-unstable and Unstable and an explanatory variable $\underline{x} = (x_1, x_2, \dots, x_5)$ the vector of explanatory variables (co-variables). The j categories of Y conditional to the values of the co-variables occur with probability $\pi_1, \pi_2, \dots, \pi_j$ that is $\pi_j = Pr(Y = j/x)$ for $j=1,2,\dots,J$. In modeling the ordinal response the cumulative probabilities $Y_{ij} = \pi_{i1} + \pi_{i2} + \dots + \pi_{ij}$ was considered. Thus this work presents ordinal logistic regression with cumulative probabilities. The explanatory variables for this work includes; current wellbeing, current concerns, concerns on mental health, mental health assessment and mental health coping strategies. The analysis was done using the R software [16]. We made use of the polr command from the MASS package to estimate the ordered logistic regression model. The command name comes from proportional odds logistic regression, highlighting the proportional odds assumption in the model. The assumption was tested using the brant command in R. The goodness of fit test was carried out to ascertain if the model was appropriate.

4.1. Model Presentation and Parameter Estimation

The cumulative logit model for this work is given by;

$$\begin{aligned} g(Y_{ij}) &= \log(Y_{ij}/1 - Y_{ij}) = \log\{[Pr(Y_{ij} \leq j)/x_i] / [1 - Pr(Y_{ij} \leq j)/x_i]\} \\ &= \log\{(\pi_{i1} + \pi_{i2} + \dots + \pi_{ij}) / (\pi_{i1} + \pi_{i2} + \dots + \pi_{ij})\} \\ &= \theta_j - x_i \beta \text{ for } j = 1, \dots, J-1 \\ g(Y_{ij}) &= \theta_j - x_i \beta \text{ for } j = 1, 2 \end{aligned} \quad (5)$$

The proportional odds model assumes that the true β values are the same in all the models, since the logit ranges from j to $j-1$ and we have the response $j=3$, which gives us two logit models. Therefore the only difference in the models is the intercept term θ_j , $j=1,2$ and it satisfies the condition $\theta_j \leq \theta_j \dots \leq \theta_{j-1}$. This means that the estimates of the two ordinal models provide just one set of β estimates. Furthermore β does not depend on j , meaning that the relationship between x and y is independent of the category. By exponentiating the β estimates relative to a given predictor ie taking e^β we obtain an estimate of the odds ratio (OR) that describes the relative odds for $y \geq 2$ for values of x_i differing by one unit. This model is usually estimated using the maximum likelihood. The maximum likelihood of θ is that value that maximizes $L(\theta)$ and it is given as:

$$L(\theta) = \prod_{i=1}^n f(x_i / \theta) \quad (7)$$

Maximizing this function in eqn(5) is equivalent to maximizing the log likelihood;

$$L(\theta) = \sum_{i=1}^n \log(f(x_i / \theta)) \quad (8)$$

To obtain the estimates of the model, eqn (6) is expanded and differentiated with respect to θ and β and setting the two equations to zero.

4.2. Goodness of Fit Test

Goodness of fit test is used to determine whether the predicted probabilities deviate from the observed in a way that the ordered logistic model cannot predict. If the resulting p-value is less than the chosen significant level, it is concluded that the predicted probabilities deviates from the observed in a way that the model could not predict. This could be as a result of; incorrect link, omitted higher-order term for variables in the model or from the omitted predictor that is not in the model. If the deviation is statistically significant, then a new link function should be sort or the terms in the model will be changed. This test was done using the chi-square test. The chi-square test is based on the difference between the observed and the expected values for each category. The chi square statistic is defined as;

$$\chi^2 = \sum_i [(o_i - E_i)^2 / E_i] \quad (9)$$

where o_i is the observed value of cases in category i , and E_i is the expected value of cases in category i .

5. Results

Table 1. The Distribution of the Ordered Multi Class of the Response

Class	N	N[%]
Stable	571	81%
Mildly-unstable	130	18%
Unstable	5	1%

The percentage of stable undergraduate is high which is good, but the concern lies in the 19% that are perceived not to be mentally stable.

The brant test was defined by Rollin Brant [3] to test if the parallel regression assumption holds. It was implemented in R under the package and function called brant, this is available on CRAN. The result of the brant test is shown below in table 2.

Table 2. Brant Test

	X2	Df	probability
Omnibus	1.862354e-05	5	1.00000
Current concerns	1.723968e-06	1	0.9989524
Current_wellbeing	5.754894e-07	1	0.9993947
Concerns on MH	2.057749e-06	1	0.9988554
MH Assessment	8.493325e-06	1	0.9976747
cop_strate	1.367451e-05	1	0.9970495

H0: Parallel Regression Assumption holds

The tables shows that the parallel regression assumption holds, because all the probabilities show none significance.

Table 3. Model Estimates

Variable	Regression Coefficient	Standard Error	t-value	p-value	Confidence Interval	
					Lower Limit	Upper Limit
Current Concerns	0.5993	0.3278	1.829	0.067	-0.04306125	1.241751
Current_wellbeing	3.5909	0.4369	8.219	0.000	2.73459772	4.447154
Concerns on MH	1.5916	0.3960	4.019	0.00005	0.81551151	2.367742
MH Assessment	3.1132	0.3688	8.440	0.00000	2.39027151	3.836092
Cop_Strate	3.5016	0.4550	7.697	0.00000	2.60990976	4.393292
Intercepts:						
stable mildly_unstable	30.7014	3.2935	9.3217	0.00000		
mildly_unstable unstable	42.1020	4.4738	9.4107	0.00000		

Residual Deviance: 222.5254

AIC: 236.5254

Table 3 above shows the estimates of the model. The table shows that all the variables are significant except the current concerns, and the confidence interval for current concerns also confirms that it is not significant.

Thus the logit model is given as;

$$\text{Logit}(y_1) = 30.7014 - [0.5993\beta_1 + 3.5909\beta_2 + 1.5916\beta_3 + 3.1132\beta_4 + 3.5016\beta_5] \quad (10)$$

$$\text{Logit}(y_1) = 42.1020 - [0.5993\beta_1 + 3.5909\beta_2 + 1.5916\beta_3 + 3.1132\beta_4 + 3.5016\beta_5] \quad (11)$$

Table 4. Odds Ratio and Confidence Interval

Variables	OR	Confidence Interval	
		Lower Limit	Upper Limit
Current concerns	1.820926	0.9578527	3.46167
Current_wellbeing	36.265822	15.4035456	85.38358
Concerns on MH	4.911733	2.2603316	10.67327
MH Assessment	22.492497	10.9164575	46.34401
Cop_strate	33.168515	13.5978238	80.90636

6. Discussion

The aim of this research work is to illustrate the use of ordered logistic regression in predicting the mental health status of undergraduate (science) students in University of Port Harcourt, based on five factors. A sample of 706 students was obtained using the stratified random sampling method. The status of the mental health of the students is an ordinal response ordered according to severity of the mental health (stable, mildly unstable and unstable). In the model five covariates are included as risk factors to mental health, namely; current wellbeing, current concerns, concerns on mental health, mental health assessment and coping strategies. The goodness of fit test was carried out and the result shows that the model was a good fit. The results of the regression in table 3 show that all considered variables except current concerns are found to be significant predictors

of mental health state; this is also evident looking at its odds compared to other factors. The brant test of proportional odds assumption shown in table 2, show that the assumption holds because the p-values indicated non-significance making an acceptance of the null hypothesis. The Omnibus value is for the whole model and the other values are for the individual coefficients. The result of the odds ratio and confidence interval is given in Table 4, the confidence interval for current concerns includes one, which tells us that current concerns is not a significant risk factor on the response or it could be said that it indicates a non-effect on the response, this also confirms what the regression results in table 3 shows. The results in table 4 also show that when the current well-being of the students increase by a unit the odds of mental health state of the student being in an unstable state or mildly-unstable state versus stable state increases by 36.27%, given that any other variable is held constant. Also when the coping strategy increases the odds also increases by 33.16%, we can also observe that increase in the mental health assessment also increases the odds of mental health in unstable or mildly unstable state versus stable state, by 22.49%. A close look at the odds ratio and the regression coefficients makes us conclude that the most influential factors of the mental health is the current well-being of the student, followed by coping strategies and lastly the mental health assessment.

REFERENCES

- [1] Agresti A. (2002). Categorical Data Analysis. Wiley, second Edition.
- [2] Agresti A. (2007). An Introduction to Categorical Data Analysis, 2nd Edn.; Wiley New York: pp 400.
- [3] Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. Biometrics 46, 1171-1178.

- [4] Chandra Sekhara O. R & Alemayehu E (2015). Ordinal Logistic regression analysis to assess the factors that affect health status of students in Ambo University: a case of natural and computational sciences college Ambo University. *International Journal of Modern Chemistry and Applied Science*, 2 (3), pp153-163.
- [5] Clareci S. C., Waleska T. C., Arminda L. S. and Mery N. S., (2008). Ordinal Logistic Regression Models: Application in quality of life Studies. <https://www.researchgate.net>.
- [6] Rune H. B. C. (2015). Analysis of Ordinal Data with Cumulative Link Models-Estimation with the R-package ordinal.
- [7] Citko D., Milewska A. J., Wasilewska J. and Kaczmarek M. (2012). Ordinal logistic regression for the Analysis of skin test reactivity to common aeroallergens. *Studies in Logic, Grammar And Rhetoric*, 29(42).
- [8] Fuller A.S (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, 38: 306-347.
- [9] McCullagh P (1980). Regression Model for Ordinal Data; *Journal of Royal Statistical Society Series B* 42(2): 109-142.
- [10] Saraceno B., Ommeren V. M, Batniji R., Cohen A., Gureje O., Mahoney J., Sridhar D.& Underhill C. (2007). Barriers to Improvement of Mental Health Services in Low-Income and Middle-Income Countries; *Global Mental Health, Series*, Vol 370 (9593), pp1164-1174.
- [11] Sumonkanti D. and Rahman R. M. (2011). Application of ordinal logistic regression analysis in determining risk factor of child malnutrition in Bangladesh. *Nutritional Journal* 2011, 10:124.
- [12] Walker SH and Duncan D. B (1967). Estimation of the Probability of an Event as Function of Several Independent Variables, *Biometrika* 54:167-168.
- [13] www.cheld.org/mental-health-care-i, Mental Health Care in Nigeria: Has the time come?
- [14] www.stats.idre.ucla.edu: Ordinal Logistic Regression in R, Institute for Digital Research and Education.
- [15] www.data.princeton.edu/wws509/notes/c6s5.html: Generalized Linear Models by German Rodrigues.
- [16] www.rdro.io>CRAN>MASS. Polr: Ordered Logistic or Probit Regression in MASS.