Evaluation of Proportional Odds and Continuation Ratio Models for Smoker in Indonesia

Rini Warti[1], Anang Kurnia[2], Kusman Sadik[2]
{riniwarti@uinjambi.ac.id[1], anangk@apps.ipb.ac.id[2], kusmansadik@gmail.com[2]}

Mathematics Education Department, UIN Sulthan Thaha Saifuddin, Jambi, 36363, Indonesia[1]Statistics Department, IPB University, Bogor, 16680, Indonesia[2]

Abstract. The polytomous model is a model used for more than two categorical response data. Some models that can use for ordinal scale responses are the Proportional Odds Model, Continuation Model, Partial Proportional Odds Model, and Adjacent Model. The Proportional Odds model has the assumption of "proportionality" or parallelity to the cumulative logit. If the parallel logits assumption not fulfilled, the alternative models that can use are Adjacent Model and Continuation-Ratio. The purpose of this study is to evaluate the proportional Odds (PO) model and Continuation-Ratio (CR) for smokers in Indonesia. The data used was taken from 2017 Indonesian Demographic and Health Survey (IDHS) by classifying smokers in ordinal categories (mild, moderate, and severe). The results show there was a violation of the assumptions in the PO Model so that the CR Model was an alternative to use. Gender is a factor that has a significant influence on all response categories. Based on the value of Goodness of fit, deviance, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Mac Fadden R2 indicate that the CR Model is better to use than the Model PO.

Keywords: Categorical Response, Continuation Ratio, Goodness of fit, Proportional Odds, Smokers.

1 Introduction

The logistic regression analysis is a way to describe the relationship between one or more independent variables and a categorical response variable. Logistic regression models are categorized according to the type of categorical response variable as follows: binary logistic regression model, multinomial logistic regression model, and ordinal logistic regression models [1]. The binary logistic regression model is used to model the binary response variables, whereas the multinomial logistic regression is a simple extension of the binary logistic regression model where the response variable has more than two unordered categories.

Response variables can also measure on an ordinal scale such as the extent of disease (normal, some, severe), liver risk levels (low, medium, and high risk), and grade of class (grade I-II-III). Outcome data collected using an ordinal scale sometimes was analyzing with a binary logistic regression by creating dichotomies among the level of the outcome variable [2],[3]. Although such approaches are not incorrect, they often result in a loss of information due to collapsing (or ignoring) some categories of the response, typically resulting in a considerable loss of statistical power [4]. Additionally, the multinomial logistic regression models often used for modeling multiple category ordinal outcomes. This analysis, however, would not take into account the ordinal nature of the result [1]. Therefore, ordinal logistic

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290483

regression models are used to model the relationship between independent variables and an ordinal response variable when the response variable has a natural ordering.

Probably the most frequently used ordinal logistic regression model is the proportional odds model of McCullagh because of the simplicity of its interpretation [5]. The continuation ratio model may be best suited in situations where the individual categories of the response variable are of intrinsic interest and are not merely arbitrary groupings of an underlying continuous variable [6]. Adjacent category models often used due to their close connection to log-linear models [7].

Although these ordinal logistic regression models have different structures and strategies for constructing logits, they share the same model assumption. The typical model assumption of these models is that the log-odds do not depend on the outcome category. This assumption is also known as the "proportionality" or "parallel logits" assumption [1]. Unfortunately, this essential and strict assumption frequently does not hold.

When the assumption of parallel logits has rejected for any of the ordinal logistic regression models, an alternative model should be considered that does not impose the constraint of parallel logits. For every ordinal logistic regression model, there are two alternative models to use in cases of non-proportional odds. Non-proportional (e.g., relaxing the "parallel logits" assumption for all of the variables in a model) and partial-proportional models (e.g., relaxing the proportionality assumption only for some) were propose for the proportional odds model and the continuation-ratio model [8]. The non-proportional and the partial-proportional versions of the adjacent category model are also feasible.

The main objectives of this research were to evaluate Proportional Odds Models (Model PO) with the Continuation Ratio Model (Model CR) in smoker cases and determine factors that significantly influence smokers.

2 Materials

The modeldiscussed in this study is the Proportional Odds Model (PO)and the Continuation Ratio Model (CR).

2.1 Assumption of Proportionality

In ordinal logistic regression, these tests are used to test the equality of various categories and decide whether the assumption is valid or not. If the assumptions do not apply, the interpretation of the results will be wrong, so the correct alternative model is needed instead of the ordinal logistic regression model. The hypothesis used to test the similarity of the coefficient of the free β_k variable in each category is stated as follows [9]:

$$H_0: \beta_{1j} = \beta_{2j} = \cdots = \beta_{(K-1)j} = \beta \qquad j = 1,2,\ldots,J$$

The test statistic used to check the proportional odds assumption is the likelihood ratio test, based on value; (Tut, 2003).

$$LR = -2\{l_{POM} - l_{NPOM}\}$$

2.2Proportional Odds Model

The cumulative logit model is the easiest model to interpret and apply. When the response variable (Y) is ordinal scale with J category, then the logit used is also multilevel value, with $J-1$ the cut-off point which is assumed through a cumulative opportunity value. The cumulative opportunity at Y is expressed as [7] :

$$P(Y \leq j) = \pi_1 + \cdots + \pi_j, \quad j = 1, \ldots, J;\ P(Y \leq 1) < P(Y \leq 2) < \cdots < P(Y \leq J) = 1$$

The Cumulative Logit Model is expressed in the form of

$$\text{logit}\{\gamma_j(\mathbf{x})\} = \log\left\{\frac{\gamma_j(\mathbf{x})}{\left(1-\gamma_j(\mathbf{x})\right)}\right\} = \theta_j - \boldsymbol{\beta}'\mathbf{x};\ j = 1, \ldots, J-1$$

Where $\gamma_j(\mathbf{x}) = P(Y \leq j|\mathbf{x})$ is a cumulative opportunity. The cumulative logit model each has its own intercept with the values $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{J-1}$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a regression parameter vector. In the PO Model, each cumulative logit has its own boundary value. The β coefficient will be the same for each response variable category [4],[10]. The Cumulative Logit model in the above equation is also called the PO Model [10].

2.3Continuation Ratio Model

When estimating conditional probabilities outside of categories and individuals reaching certain categories, namely $\pi(Y > j|Y \geq j|)$, the CR model can be expressed in this form [7],[12] :

$$\ln\left(\frac{\pi(Y \geq j|x_1, x_2, \cdots, x_p)}{\pi(Y = j|x_1, x_2, \cdots, x_p)}\right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where $\pi(Y \geq j|x_1, x_2, \cdots, x_p)$ is a conditional probability outside category j, conditional in a category with a series of predictors. $j = 1, 2, \cdots, J-1$. α_j is the intersection, and $\beta_1, \beta_2, \cdots, \beta_p$ is the logit coefficient [11], [12].

2.4Odds Ratio

The odds rasio is used to interpret the proportional odds model (1), which is comparison of the cumulative opportunities with the complement. The value of the odds rasio of the event $Y \leq j$ when $X = x_1$ and $X = x_2$[5], is

$$\frac{\gamma_j(x_1)/\left(1-\gamma_j(x_1)\right)}{\gamma_j(x_2)/\left(1-\gamma_j(x_2)\right)} = \exp\{-\beta^T(x_1 - x_2)\}$$

The odds ratio of the cumulative opportunity is called the cumulative odds ratio. The log value of the cumulative odds ratio is proportional to the distance between and. In other words, the correlation between independent variables and response variables does not change for each category of response variables, and parameter estimates do not change for each cut-off point [9].

3Methods

The research data obtained from the Indonesian Demographic and Health Survey (SDKI), 2017, regarding smokers with the number of observations in this study, was 16,336 respondents. The variables observed in this study were many cigarettes consisting of 3 categories (polytomous). The categorization of smokers in this research based on the number of cigarettes was smoke per day. Which used as a variable polytomous response. The explanatory variables used in forming the polytomous model explained in Table 1.

Table 1 Category of Response Variables and Explanatory Variables

Variable	Scale	Category		
Many cigarettes (Y)	Ordinal	1-10 stem per day	Light smoker	1
		11-24 stem per day	Moderate smoker	2
		> 24 stem per day	Heavy smoker	3
Gender (X1)	Nominal	Male		0
		Female		1
Age (X2)	Ordinal	Teenager		0
		Adult		1
Status (X3)	Nominal	Married		0
		Single		1
		Others		2
Residence (X4)	Nominal	Urban		0
		Rural		1
Education (X5)	Ordinal	No education		0
		Elementary school		1
		Junior high school		2
		Senior high school		3
		College		4
Occupation (X6)	Nominal	Agriculture		0
		Industry field		1
		Service sector		2
		Others		3

The primary response variable in the form of nominal scale data that converted into categorical data, commonly known as ordinal variables ordered from light smokers, moderate smokers, and heavy smokers. Then the data will be analyzed using the PO Model and CR Model with the following stages:

1. Detect the initial fulfillment of the proportional odds assumptions of the PO Model
2. Establishing a PO Model and Model CR using the VGAM package.
3. Test the hypothesis of proportionality from the cumulative logit by using the likelihood ratio test.
4. Compare the two models to choose the best model through the goodness of fit indicator.
5. Determine factors that significantly affect smokers in Indonesia based on the best-chosen model.
6. Analyze the odds ratio of the factors that significantly influence the model.

4Results and Discussion

The model discussed in this study is the Proportional Odds Model and the Continuation Ratio Model (CR).

4.1 Description of Data

The results of the 2018 Ministry of Health's Basic Health Research showed an increase in the prevalence of tobacco consumption in people aged > 15 years from 32.8% (in 2016) to 33.8% (in 2018) with 12.3 cigarettes smoked per day.

4.2Proportional Odds (PO) and Continuation Ratio (CR) Model

Proportional odds assumptions must fulfill in building the PO model; namely there are equal proportions in each response variable category. The parameter similarity of the six explanatory variables was tested using the likelihood ratio test. The test results show that the likelihood ratio value is 12.025, and the P-value is 2.2e-16, so it can conclude that the assumption of proportional odds violated or not fulfilled.

Before the research conducted, it determined that the real level (α) in this study was 5%. The estimation results of the Proportional Odds Model (PO) with the smoker category as the response variable and the five explanatory variables shown in Table 2

Table 2 Results of Analysis of the Proportional Odds Model

Variables	Estimation	p-value
(Intersep):1	-2.8571	< 2e-16
(Intersep):2	-0.6967	0.005
X1(1)	0.9217*	< 2e-16
X2(1)	0.3968	0.2390
X3(1)	0.2812	0.2159
X3(2)	0.3110	0.2722
X4(1)	0.0108	0.8457
X5(1)	0.0064	0.9720
X5(2)	-0.0759	0.4288
X5(3)	-0.0862	0.3306
X6(1)	-0.0299	0.6747
X6(2)	-0.0839	0.2658
X6(3)	-0.0782	0.3209

The results of the analysis of the Proportional Odds Model (PO) in Table 3 show that only one significant variable in the model is X1 (gender). But this result cannot be concluded because it will provide invalid conclusions so that the analysis phase needs to be done using an alternative model, namely the Continuation Ratio (CR) model. Parameter estimation results using the Continuation Ratio (CR) model, as shown in Table 3.

Tabel 3 Results of Analysis of the Continuation Ratio (CR) Model

Variables	Estimation	p-value
(Intersep):1	-2.8588	< 2e-16
(Intersep):2	-0.9085	7.58e-05
X1(1)	0.9260*	< 2e-16
X2(1)	0.3929	0.198
X3(1)	0.2902	0.165
X3(2)	0.3287	0.208
X4(1)	0.0122	0.816
X5(1)	0.0108	0.950
X5(2)	-0.0672	0.459
X5(3)	-0.0801	0.340
X6(1)	-0.0289	0.669
X6(2)	-0.0894	0.211
X6(3)	0.0734	0.325

The results of the analysis of the Continuation Ratio (CR) model in Table 4 show that only one significant variable in the model is X1 (gender). This result is the same as using the PO Model, meaning that only gender variables have a real influence on the smoker category in Indonesia.

4.3Model Feasibility Test

The model feasibility indicator uses four values, namely Deviance, AIC, BIC, and R2 McFadden, as shown in Table 4.

Table 4 Model Feasibility Test

Model Feasibility Indicator	PO Model	CR Model
Deviance	11401.12	11389.09
p-value	0.000	0.605
AIC	11427.12	11415.1
BIC	11515.66	11503.64
Mc Fadden R^2	0.321	0.591

From Table 4, it can see that the CR Model is better than the PO model based on these four indicators. The feasibility test model based on deviance value states that the CR model significantly (P-value = 0.605) is said to be more feasible to use compared to the Model PO. Both AIC and BIC values for the CR model are also smaller than those of the PO model. The goodness of the CR model can explain smokers is around 59.1% based on the value of R2 Mc-Fadden compared to the value of R2 Mc Fadden for the PO model which is only able to explain 32.1%.

5Conclusion

The results of this study indicate that if not fulfilled the proportional odds assumptions in the PO Model, the alternative model that can use is the CR Model. The CR model with five

explanatory variables able to explain smoker models in Indonesia better than the PO Model, with a value of 59.1%. The tendency of mild smokers compared to heavy smokers occurred in male sex compared to women with an odds ratio of 2.51.

The CR model is better than the PO Model with a lower value of the goodness of fit of AIC and BIC. The results of the CR Model analysis indicate that there are still several categories of explanatory variables that have no significant effect. Therefore several alternative models can be used in further research. Alternative models suggested to analyzed in this case include the Non-Proportional Odds Model, the Partial Proportional Odds Model, and the Adjacent Category Model.

References

[1] Hosmer, D. W., Lemeshow, S.:Applied Logistic Regression.(2th edn). Wiley. New York.(2000)

[2] Bender, R., Grouven, U.: Using Binary Logistic Regression Models For Ordinal Data With Non-Proportional Odds. Journal of Clinical Epidemiology.Vol. 51, pp. 809–816.(1998)

[3] Norris, C. M. et al.: Ordinal Regression Model and The Linear Regression Model were Superior to The Logistic Regression Models.Journal of Clinical Epidemiology.Vol. 59, pp. 448–456.(2006).

[4] Ananth, C. V., Kleinbaum, D. G.: Regression Models for Ordinal Responses: A Review of Methods and Applications.International Journal of Epidemiology.Vol. 26, pp. 1323–1333.(1997).

[5] McCullagh, P.: Regression Models for Ordinal Data (with discussion). Journal of the Royal Statistical Society Series B. Vol. 42, pp. 109–142.(1980).

[6] Greenwood, C.; Farewell, V.: A Comparison of Regression Models for Ordinal Data in An Analysis of Transplanted-Kidney Function. The Canadian Journal of Statistics.Vol. 16, pp. 325–335.(1988)

[7] Agresti Alan. An Introduction to Categorical Data Analysis (Second Edition). John Wiley & Sons, Inc. Canada.(2007)

[8] Dolgun A, Saracbasi O: Assessing Proportionality assumption in the adjacent category logistic regression model. Statistics and Its Interface.Vol. 7, pp. 275-295.(2014)

[9] Ari E, Yildiz Z.: Parallel Lines Assumption In Ordinal Logistic Regression And Analysis Approaches. International Interdisciplinary Journal of Scientific Research ISSN: 2200-9833, pp. 8-23. (2014).

[10] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Second Edition. Chapman and Hall. London. 511 p. (1989).

[11] Liu, X.; O'Connell A, Koirala, H.: Ordinal Regression Analysis: Predicting Mathematics Proficiency Using the Continuation Ratio Model. Journal of Modern Applied Statistical Methods. Vol. 10, No. 2, pp. 513-527.(2011).

[12] Liu, X: Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata. NERA Conference Proceedings. Eastern Connecticut State University.(2010)