The comparison of the ordinal logistic model with the classical regression model

N A Yensy*
Program Studi Pendidikan Matematika, Universitas Bengkulu, Jl. WR Supratman, Kota Bengkulu 38122, Indonesia

* nurulastutyyensy@unib.ac.id

Abstract. This research’s aims was to provide an alternative for statistical user who were interested in analysing the functional relationship between ordinal response and explanatory variable. The data used was the data of Students on Generation 2017 JPMIPA FKIP Bengkulu University. The test statistic from the modified probability ratio in ordinal logistic had a higher test power than the F test in the classical linear regression. Overall the probability of rejection of ordinal logistic tended to be lower than the classical regression. At condition of ρ close to one, the test power of the ordinal logistic would be the same as the classical regression. Ordinal logistic model tended to be more selective in choosing variables that significantly affect the response, resulting in a simpler model than the classical regression. The results of the research with the classical linear model obtained variables that influence the student GPA were UN score, High School Report, UNIB admission pathway, High School Status and interaction of UN Score and High School Status. While the ordinal logistic model produced influential independent variables were the UN Score, High School Report, High School Status and the interaction of UN Score and High School Status.

1. Introduction
In statistical modeling, one important thing to note is the response variable's data structure, without ignoring the role of the independent variable [1,2]. The response variable can be a nominal, ordinal, interval or ratio scale that has a tiered nature, namely the lowest nominal scale and the highest ratio scale. This means that the principle applies that the statistical method applied for a particular scale can be used for a higher scale, otherwise it cannot be used for a lower scale [3]. Furthermore, the statistical modeling on continuous scale response variable (ratio or interval) would be good enough if approached with classical regression with the assumption of normality, freedom and homogeneity of variances. However, the response variable is often found in a categorical structure [4,5].

The ordinal response data are usually more often analyzed by using a cross tabulation that are actually applied to nominal response. Although this does not violate the principle of the statistic, this method has several weaknesses, including the nature of the settings owned by the ordinal scale is not heeded, the change in response due to change in the independent variable cannot be calculated and if there are independent variable measured on a scale above nominal then the information about the independent variable will also be ignored [6-8]. To overcome this limitation, a multinomial logistic model has been developed which is an extension of the binary logit model in which the classification of response is expanded into three or more categories [9]. As for the ordinal response data, statistical modeling can be approached with an ordinal logistic model which is part of a multinomial logistic model [10].
The logistic model coefficients are easier to interpret than the other models [11]. The study of ordinal logistic model is carried out by tracing the underlying theory and their application. The search of theory is done with the aim to obtain a coherent translation in order to obtain a complete picture of both the usefulness and relationship with the other statistical method. Many the previous research have dealt with ordinal logistic regression, but only limited to the application of statistical models. It has never been tested or compared the goodness of the model's fit with the other statistical models. This research will show that the ability of the ordinal logistic model is not much different from the classical regression by looking at the power of the test from the test statistic used in both models. As a practical approach the data of the academic records of students on generation 2017 JPMIPA FKIP University of Bengkulu is sought to find out what the factors that influence the achievement of students' academic achievement during lectures in the first year.

The purpose of the research was to compare the power of F test in the classical regression model with the probability ratio test statistic (G) in the ordinal logistic model and compare the result that obtained in the ordinal logistic model with the classic linear regression model.

1.1. Logistic regression

Binary logistic regression is a data analysis method used for searching the relationship between the response variable (Y) which is binary or dichotomous with the predictor variable (X) which is polycotomous [12]. The output of the Y consists of two categories usually denoted as Y = 1 (success) and Y = 0 (failed). The general form of logistic regression is as follows:

$$ \Pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} $$

(1)

Then the logit transformation is performed:

$$ g(x) = \ln \left[ \frac{\Pi(x)}{1-\Pi(x)} \right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k $$

(2)

If there are j categories of response, the ordinal logistic model is formed:

$$ Y_j(X) = \Pi_1(X) + \Pi_2(X) + \cdots + \Pi_j(X) $$

(3)

2. Methods

2.1. Simulation method

To see the test power of the classical linear regression and the ordinal logistic regression simulation is performed by using MINITAB 16.0 software, to see the magnitude of the rejection proportion of H₀: β = 0. The procedure used in this simulation is as follow:

Initially generated n paired data consisting of one dependent variable y and k the independent variable x, x = (x₁, x₂, ..., xₖ). All variables raised are from the population that have a standard normal distribution and are uncorrelated, except for the correlation between xᵢ and y of ρᵢ for all i = 1, 2, ..., k. In the classic linear regression model, the testing procedure to determine the accept or the reject of H₀ used is the F test against the ratio of the middle square of the regression and the middle square of the error, with the F distribution and free degree k and n - k - 1 [13,14].

Whereas to apply the use of the ordinal logistic regression model, the values of the variable y are transformed in the form of quartile interval, so that they are divided into four categories. The testing procedure used is the modified F test which is the result of the work of Snappin and Small (1996). In this procedure, that is tested the probability ratio test (G) and the modified probability ratio test (M). In this simulation the sample sizes tested were 10, 25 and 50, while the ρ values were 0.0, 0.2, 0.4 and 0.6. Meanwhile, the number of explanatory variables that were tried were 1, 3 and 5. From each sample
generated, it was passed through two modeling procedures and tested at the significant level of 0.01, 0.05 and 0.10. From each combination the number of independent variables (k) and the number of examples (n) will repeated 1000 times, then recorded the proportion of rejection $H_0$: $\beta = 0$ at various of values $p$.

2.2. Practical approach method

The data used in this research was the data of students on generation 2017 JPMIPA FKIP Bengkulu University. Data was obtained from the secretariat of the academic section. The data required was a student's GPA for two semesters (semester I and II), which were as the response variables with 5 value categories, which were as follow:

- GPA < 2.0 worth 1;
- 2.0 ≤ GPA < 2.5 worth 2;
- 2.5 ≤ GPA < 3.0 worth 3;
- 3.0 ≤ GPA <3.5 worth 4; and
- GPA ≥ 3.5 worth 5

Furthermore, the independent variables which are estimated to be able to explain the students' academic achievement were presented in table 1. Data with continuous response (GPA) was modeled with the classical linear regression.

**Table 1. Kind of independent variable.**

| Description                              | Variable Type | Notation |
|------------------------------------------|---------------|----------|
| Gender                                   | Binary        | $X_1$    |
| 0 = female, 1 = male                     |               |          |
| Father's education                       | Category      | $X_2$    |
| 1 = Basic education                      |               |          |
| 2 = Secondary education                  |               |          |
| 3 = Higher education                     |               |          |
| Father's occupation                      | Category      | $X_3$    |
| 1 = PNS, 2 = ABRI, 3 Retired of PNS/ABRI |               |          |
| 4 = Self-employed,                       |               |          |
| 5 = private /BUMN/ Professional/ executive employee, |   |          |
| 6 = farmer/fisherman, 7 =other things    |               |          |
| Parents' income (in Rupiah)              | Continuous    | $X_4$    |
| 1 = Basic education                      | Category      | $X_5$    |
| 2 = Secondary education                  |               |          |
| 3 = Higher education                     |               |          |
| Mother's occupation                      | Category      | $X_6$    |
| 1 = PNS, 2 = ABRI, 3 Retired of PNS/ABRI |               |          |
| 4 = Self-employed, 5 = private /BUMN/ professional/ executive employee, |  |          |
| 6 = farmer/fisherman, 7 =other things    |               |          |
| High School Status                       | Binary        | $X_7$    |
| 0 = public, 1 = private                  |               |          |
| Average of High School Report            | Continuous    | $X_8$    |
| UN Score                                 | Continuous    | $X_9$    |
| UNIB admission pathway                   | Binary        | $X_{10}$ |
| 0 = SPMU, 1 = SNMPTN                     |               |          |
| Suitability of study program             | Binary        | $X_{11}$ |
| 1 = in accordance with the first choice  |               |          |
| 0 = does not match with the first choice |               |          |
3. Results and discussion

3.1. Simulation result

Based on the simulation that have been done, it is obtained that in general the level of significance obtained from the modified probability ratio test (M) tends to be closer to the specified significant level (α) than the standard ratio test (G), while comparing the test power of the two tests will valid when the same real level is used. As such, it is feasible to compare the modified ratio test statistic (M) with the F test statistic in classical regression. The modified ratio test statistic (M) is assumed to be distributed in F with free degree of k and n - k - 1.

The use of modified ratio test in the ordinal logistic generally has a lower chance of rejection of \( H_0: \beta = 0 \) than the use of standard F test in the classical linear regression. This situation will make the test power of the M test statistic on the ordinal logistic always be higher than the classical regression at the \( \rho \) condition close to zero. Likewise, when \( \rho \) approach of 1, the proportion of rejection \( H_0: \beta = 0 \) in the ordinal logistic is not much different from the classical regression. The slight difference is only at the significance level \( \alpha = 0.01 \) with a sample that is not too large \( (n = 10) \), where the selection of level of \( \alpha = 0.01 \) is very strict. In general, the ordinal logistic model will have a high chance of rejecting \( H_0: \beta = 0 \) if the independent variable is really closely related to the response variable. This will make the ordinal logistic model more selective than the classical regression in choosing the independent variables to be the variable that have a significant effect on the response. This more selective nature will produce a simpler model but can still explain the state of the data. That is, the statistical model that is formed will contain the fewest variables by providing the same information as the model for more variables. These results are in accordance with the theory by Snappin and Small [15] namely about the test of significance using regression model for ordered categorical data, He said that the test power of the modified probability ratio tests on the ordinal logistic model always be higher than the classical regression model.

3.2. Practical approach result

The modeling begins with determining which explanatory variables that significantly affect the GPA if included in the model, both in the linear regression and the ordinal logistic regression. After being analyzed, it was found that UN score and Report were quite influential on the acquisition of GPA, with a correlation value (\( \rho \)) = 0.65 for the UN value with GPA. Furthermore, UN scores with Report (\( \rho = 0.649 \)) and Report with GPA (\( \rho = 0.497 \)) for a combination of private and state high school. Whereas for state high school, the correlation of UN scores with GPA is \( \rho = 0.646 \), the correlation of UN scores with Report is \( \rho = 0.651 \) and the correlation of Report with GPA is \( \rho = 0.495 \). Whereas for private high school, the correlation of UN scores with GPA is \( \rho = 0.675 \), the correlation of UN values with Report is \( \rho = 0.647 \) and the correlation of Report with GPA is \( \rho = 0.516 \).

Furthermore, it is also explored the possibility of interaction between explanatory variable, especially the interaction between explanatory variable that are interval scale and dummy explanatory variable. From these result it was found that the UN scores interacted with High School Status and the suitability of choice in influencing the magnitude of the GPA. Furthermore, along with the real independent variables in the single variable regression, interaction variables are entered into the model, both in the classical regression model and logistic regression.
Table 2. The estimated value of the coefficient on the multiple linear regression model and its partial test.

| Independent Variable                      | Estimated Coefficient | Standart deviation Coefficient | T_count | p-value |
|------------------------------------------|-----------------------|--------------------------------|---------|---------|
| Constants                                 | -0.43920              | 0.20010                        | -2.19   | 0.028   |
| Gender                                    | -0.03215              | 0.01958                        | -1.64   | 0.101   |
| High School Status                        | -1.04270              | 0.22290                        | -4.68   | 0.000   |
| UN score                                  | 0.05464               | 0.00272                        | 20.04   | 0.000   |
| Pathway                                   | -0.18254              | 0.02343                        | -7.79   | 0.000   |
| REPORT                                    | 0.08161               | 0.03055                        | 2.67    | 0.008   |
| Suitability                               | -0.23340              | 0.14380                        | -1.62   | 0.105   |
| UN score*High school status               | 0.02385               | 0.00508                        | 4.69    | 0.000   |
| UN score * pathway                        | 0.00532               | 0.00322                        | 1.66    | 0.098   |

Table 2 shows that partially, the independent variables that significantly affected the response were High School Status, UN scores, Pathway, Reports and interaction between UN scores and High school status at a significant level of 5% due to p-values < 0.05. While the other explanatory variables were not significant.

Further more, the coefficient of determination ($R^2$) generated by involving the independent variable as in table 2 is 45.3%. Whereas if the independent variable in the above model whose the partial test has a p-value greater than 0.05 are not included in the model, an $R^2$ value of 45.6% is obtained.

Table 3. The variance of multiple linear regression model.

| Source of variance | Free degree | Number of square | Middle Square | $F_{count}$ | p-value |
|--------------------|-------------|------------------|---------------|-------------|---------|
| Regression         | 8           | 380.175          | 47.522        | 228.8       | 0.000   |
| Error              | 186         | 38.688           | 0.208         |             |         |
| Total              | 194         | 418.863          |               |             |         |

Table 3 shows that simultaneously, the explanatory variables significantly influence the response (GPA) because the $F_{count} < 0.05$. Next, the estimated coefficient and variance for the new model can be seen in table 4 and 5 below:

Table 4. The estimated value of the coefficient of the multiple linear regression model and its partial test if the variable that not significantly removed from the model.

| Independent Variable                      | Estimated Coefficient | Standart deviation Coefficient | T_count | p-value |
|------------------------------------------|-----------------------|--------------------------------|---------|---------|
| Constants                                 | -0.576                | 0.1862                         | -3.10   | 0.002   |
| High School Status                        | -1.050                | 0.2229                         | -4.71   | 0.000   |
| UN score                                  | 0.05725               | 0.00221                        | 25.89   | 0.000   |
| Pathway                                   | -0.1854               | 0.02339                        | -7.93   | 0.000   |
| REPORT                                    | 0.0829                | 0.03052                        | 2.72    | 0.007   |
| UN score * High school status             | 0.02394               | 0.00508                        | 4.71    | 0.000   |

Table 4 shows that after the explanatory variables which did not significantly affect the response were removed from the model, the estimated coefficient of regression was obtained and that all independent variables were significant.

Then, the student of the state high school tend to have higher GPA, the student who entered UNIB through the SNMPTN pathway turned out to have a smaller GPA than the student from the SPMU
pathway. There is a tendency for the student who have a high average of UN score to be followed by a high GPA. This tendency mainly occurred in the student who were in private high school. Likewise with the REPORT score, the student that have a high REPORT grade will be followed by a high GPA.

Table 5. The final variance of the classic multiple linear regression model.

| Source of variance | Free degree | Number of square | Middle Square | $F_{count}$ | $p$-value |
|-------------------|-------------|-----------------|--------------|-------------|-----------|
| Regression        | 5           | 379.041         | 75.808       | 364.71      | 0.000     |
| Error             | 189         | 39.312          | 0.208        |             |           |
| Total             | 194         | 418.353         |              |             |           |

Table 5 shows that simultaneously, the explanatory variables significantly influence the response (GPA) because the $F_{count} < 0.05$.

Furthermore, the examination of assumption of residual for the purpose of testing the hypothesis show that there were no significant deviation from the behavior of the residual, both in terms of normality, randomness and homogeneity of variance. By fulfilling this assumption, then testing on this regression model can be considered valid.

In the ordinal logistic regression model, the estimated coefficient obtained through the maximum likelihood method as well as the partial test result used the Wald Test as in table 6 below:

Table 6. The estimated value of the coefficient of the ordinal logistic model, partial test and odds ratio value.

| Independent Variable    | Estimated Coefficient | Standart deviation Coefficient | Z-Wald | $p$-value | Odds-Ratio |
|-------------------------|-----------------------|--------------------------------|--------|----------|------------|
| Constants (1)           | 10.0582               | 0.85300                        | 11.79  | 0.000    |            |
| Constants (2)           | 12.2169               | 0.86470                        | 14.13  | 0.000    |            |
| Constants (3)           | 13.8853               | 0.87600                        | 15.85  | 0.000    |            |
| Constants (4)           | 15.7478               | 0.88910                        | 17.71  | 0.000    |            |
| High School Status      | 0.6827                | 0.09731                        | 7.02   | 0.000    | 1.98       |
| UN Score                | -0.2089               | 0.01225                        | -17.05 | 0.000    | 0.81       |
| Pathway                 | 1.1296                | 0.63050                        | 1.79   | 0.073    | 3.09       |
| Report                  | -0.4250               | 0.12590                        | -3.37  | 0.001    | 0.65       |
| Suitability             | 4.4540                | 1.18000                        | 3.77   | 0.000    | 86.01      |
| UN Score * High School Status | -0.0993           | 0.02634                        | -3.77  | 0.000    | 0.91       |
| UN Score * Suitability  | -0.0257               | 0.01400                        | -1.83  | 0.067    | 0.97       |

Table 6 shows that the explanatory variable of pathway and the interaction of variable of UN score with suitability were not significant based on the $p$-value. The model had a log-likelihood value of -2716.538 with a log-likelihood model without explanatory variable of -3340.421, statistic value G = 1224.240 and $p$-value = 0.000, which mean that the existence of the explanatory variable was able to explain the GPA quite well. By using Pearson and Deviance statistic, the goodness of fit test showed that the rest of the model can be tolerated, is considered quite small, because the $p$-test values are respectively 0.994 and 1.00.

The estimated value of the coefficient of PATHWAY variable showed that the student from SPMU pathway tend to have a better IPK. Likewise, the alleged value for the high school STATUS variable, which identified that the higher GPA of the student that coming from state high schools compared to the student from private high schools. The level of achievement when in high school, which is indicated by UN Score and Report grade had a positive association with the GPA obtained. The suitability of the
choice of majors at UNIB is also an important indicator for the acquisition of student’s GPA, this was evident from the high enough OR score (OR = 86.01) which showed that the student with a choice of majors by UNIB according to personal choice (choice I) tend to get a higher GPA better than the student that had a choice not in accordance with the first choice.

Based on the best logistic model with α = 0.05, it can be concluded that the variable that were estimated to influence the acquisition of GPA were UN score, Report, High School Status and interaction of UN Score and High School Status.

3.3. Comparison of both models
In terms of interpretation based on the estimated value of the explanatory variable coefficient in the model, there were no deviation between the two models obtained. The direction of the effect of each the explanatory variable on multiple linear regression with the least squares method and the ordinal logistic regression also did not differ. Furthermore, the partial p-test value by using the t test in the classical linear regression and the Wald test in the logistic regression for each of the alleged explanatory coefficient were almost the same. This fact showed that the changing of measurement scale by making the GPA in the categorical-ordinal form did not result in meaningful changes in the interpretation of the results obtained.

The result obtained in the logistic regression model were not much different from this classic regression also due to the division of the GPA response into 5 categories quite close to the normal distribution. However, through the logistic model it could be clearly seen the difference in response opportunities due to the Odds Ratio value. Thus, for this case the ordinal logistic model was better than the classic linear regression, specifically for the purpose of classification and comparison of the achievement of each category of explanatory variable that affect the response (GPA). The logistics model was also quite good for the purpose of classifying response variable [16,17].

4. Conclusion
The test statistic from the modified probability ratio in the ordinal logistic had a higher test power than the F test in the classic linear regression for H0: β = 0. Overall the probability of rejection (H0: β = 0) of ordinal logistic tends to be lower than the classical regression. At condition ρ close to 1, the test power of the ordinal logistic will be the same as the classical regression. The ordinal logistic model tend to be more selective in choosing variable that significantly affect the response, resulting in a simpler model than the classical regression.

Based on the practical approach, it can be concluded that the variables that influence the acquisition of GPA according to the classical linear model with a significant level α = 0.05 were UN score, REPORT, pathway, High School Status and interaction of UN score and High School Status. While the ordinal logistic model with the same significant level produced the influential independent variables were UN scores, REPORT, High School Status, Suitability of Study Program and interaction of UN Score and High School Status.

Acknowledgments
We are grateful to academic staff of FKIP, Bengkulu University, for their help and support for providing the students’ academic data for this research.

References
[1] Fahrmeir L and Gerhard T 2004 Multivariate Statistical Modelling Based on Generalized Linear Models (New York: Springer-Verlag) pp 67-85
[2] Cullagh P and Jhon A N 1993 Generalized Linear Models 2nd Ed (New York: Chapman and Hall) pp 123-147
[3] Aitkin M, Anderson and Hinde J 1999 Statistical Modelling in GLIM (Oxford: Clorendon Press) pp 59-71
[4] Agresti A 1994 Analysis of Ordinal Categorical Data (New York: John Wiley and Sons) p 54
[5] Hosmer D W and Lemeshow 1999 *Applied Logistic Regression* (USA: John Wiley and Sons) p 95

[6] Thomas P, Chandra A, Bhatia T, Mishra N N, Sharma V R, Gauba D ... and Nimgaonkar V L 2011 Clinical and genetic correlates of severity in schizophrenia in India: an ordinal logistic regression approach *Psychiatry research* **189**(2) 321-323

[7] Ezgi A D, Sermet A and Gulser K 2009 Determination of optimal product styles by ordinal logistic regression versus conjoint analysis for kitchen faucets *International Journal of Industrial Ergonomics* **39** 866–875

[8] Collet D 2001 *Modelling Binary Data* (London: Chapman and Hall) pp 73-79

[9] Made A D 2015 Application of ordinal logistic regression for analyzing level of victims of accident traffic district Buleleng *E-Journal of Mathematics* **4** 54-58

[10] Frangos C C, Frangos C C and Sotiropoulos I 2011 Problematic internet use among Greek university students: an ordinal logistic regression with risk factors of negative psychological beliefs, pornographic sites, and online games *Cyberpsychology, Behavior, and Social Networking* **14**(1-2) 51-58

[11] Ninda W P, Yohanes A R and Jantje D 2016 Ordinal Logistics Regression Analysis in the Service User Statisfaction Level Toward Service at Sam Ratulangi International Airport Of Manado *E-Journal of Mathematics* **5** 72-79

[12] Baker R D, Weinand C, Jeng J C, Hoeksema H, Monstrey S, Pape S A ... and Wilson D 2009 Using ordinal logistic regression to evaluate the performance of laser-Doppler predictions of burn-healing time *BMC medical research methodology* **9**(1) 11

[13] Mwalili M S, Lesaffre E and Declerck D 2005 A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study *Journal of Application Statistics* **54** 77–93

[14] Imaslihkah S, Ratna M and Ratnasari V 2013 Ordinal Logistic Regression Analysis of Factors that Affect the Predicate Graduation of Students at ITS Surabaya *Journal of Science and Arts Pomits* **2** 177-182

[15] Snappin M S and Small D 1996 Test of Significance Using Regression Models for Ordered Categorical Data *Biometrics* **42** 583-592

[16] Fuks M and Salazar E 2008 Applying models for ordinal logistic regression to the analysis of household electricity consumption classes in Rio de Janeiro, Brazil *International Journal of Energy Economics* **30** 1672–169

[17] Parsons N R, Costa M L, Achten J and Stallard N 2009 Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R *Computational Statistics and Data Analysis* **53**(3) 632-641