Handling Overdispersion Problems in Multinomial Logistic Regression (Study Case in Stress Level Data)

A'YUNIN SOFRO[1,*], KHUSNIA NURUL KHIKMAH[2], DANANG ARIYANTO[3], YUSUF FUAD[4], BUDI RAHADJENG[5], YULIANI PUJI ASTUTI[6]
[1-6]Mathematics Department
Universitas Negeri Surabaya
Ketintang Street, Surabaya, 60231, East Java
INDONESIA

Abstract: The development of statistical methods also impacts the development of analytical methods. One analytical method in which this is the case is the multinomial logistic regression modeling method. In this method, we have more than two categories of the response variable. At this time, the data used in modeling has various problems, one of which is overdispersion. This is a condition where there is a correlation between the response variables. This paper will examine the performance of multinomial logistic regression when there is overdispersion present in the data. We will focus on implementing methods in the Stress Level Data, which is about student stress level due to 'zoom fatigue'. The model selection is carried out using the stepwise method, where the best model is selected based on the smallest AIC value of the model candidates. The best model for our data shows that the performance of the multinomial logistic regression approach with overdispersion treatment is better than without allowing for overdispersion.

Key-Words: Multinomial Logistic Regression, Overdispersion, Stress Level Data, Zoom Fatigue

Received: March 8, 2023. Revised: October 25, 2023. Accepted: November 18, 2023. Published: December 31, 2023.

1 Introduction

This has given rise to many modeling techniques. Statistical modeling is one method that aims to find information from the data that will be used for various purposes in various fields [1]. The data available in everyday life also has various types, one of which is discrete data. This type of discrete data in modeling requires a variety of approach methods, one of which is multinomial logistic regression [2].

Multinomial logistic regression is a statistical method in which the response variable has more than two categories. This method is often referred to as a development of the binomial logistic regression method [3]. Determine the relationship between the response variable and several explanatory variables. However, in practice, the data can present several difficulties. One of them is overdispersion.

Overdispersion is a condition where there is a correlation between the response variables [4]. If the data show an overdispersion, the impact of the result is poor modeling [5]. Therefore, it is necessary to deal with the problem of overdispersion. Goodness-of-fits test can evaluate the performance of modeling. Therefore, it is necessary to calculate the goodness of the modeling method's goodness before and after handling overdispersion, and this is one of the aims.

In our study, the data with this overdispersion problem is found in the Stress Level Data, which is about the stress level in students due to zoom fatigue during online lectures. If it is not treated, zoom fatigue itself will have various negative impacts on the body. For example, the problem of fatigue [6], lack of socialization [7], and mental health [8] cause potential stress [9]. This problem has a severe impact and so it needs further analysis.

Therefore, this paper will focus on handling the overdispersion of data show subject to multinomial logistic regression. Our methods will apply to stress level data. The aim is to explore potential factors that impact stress due to zoom fatigue.

2 Materials and Methods

2.1 Data Sources

The data used in this article is Stress Level Data, which are secondary data taken from Kaggle (www.kaggle.com/datasets/iganarendra/zoom-

fatigue-pada-mahasiswa-indonesia) last accesses at. More briefly, the explanatory variables used in this paper has shown in Table 1 below.

Table 1. The Explanatory Variables

Variable	Description	Scale
X_1	Average usage	Ratio
X_2	Gender	Nominal
X_3	Device Type	Nominal
X_4	Screen size	Ratio
X_5	With Desk	Ordinal
X_6	With sofa	Ordinal
X_7	On the floor	Ordinal
X_8	Lying	Ordinal
X_9	Mobile	Ordinal
X_{10}	Audio type	Nominal
X_{11}	Usage limit	Ratio
X_{12}	Eye condition	Ordinal
X_{13}	Waist condition	Ordinal
X_{14}	Tingling	Ordinal
X_{15}	Neck condition	Ordinal
X_{16}	Head condition	Ordinal
X_{17}	Finger condition	Ordinal

The research response variables are more frequent. It is presented in Table 2 below.

Table 2. Response variables

Category	Description	Number of students
1	No stress (Normal)	115
2	Experiencing Mild Stress	151
3	Experiencing Moderate Stress	20
4	Experiencing Severe Stress	2

This study will model the data using the multinomial logistic regression method. Then it will check whether there is overdispersion in the data or not. Suppose the data are affected by overdispersion. In that case, the data will again be modeled with multinomial logistic regression with overdispersion.

The first data exploration is the descriptive statistics steps for this study's analysis. Then we test the independence of the explanatory variables on the response variables to see whether they are correlated. After that, we analyze the data with multinomial logistics regression without and with the overdispersion case, and evaluate the performances.

2.2 Multinomial Logistic Regression

Multinomial logistic regression is one of the data analysis methods in statistics that aims to find the relationship between the response variable (Y) and several explanatory variables (X). The response variable (Y) is multinomial or has more than two categories. Suppose, $i = 1,2,\dots,n$, where n is the number of observation, where $\sum_{j=1}^{p} \pi_j = 1$. In general the multinomial logistic regression model can be written as follows [10]:

$$\pi_j(x) = P(Y = j|x) = \frac{e^{(\beta_{j0}+\beta_{j1}x_1+\beta_{j2}x_2+\cdots+\beta_{jk}x_k)}}{1+e^{(\beta_{j0}+\beta_{j1}x_1+\beta_{j2}x_2+\cdots+\beta_{jk}x_k)}} \quad (1)$$

where $j = 1,2,3,\dots,p$, the number of outcome in Y is defined as p and k is the number of explanatory variable. Let $g_j(x) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jk}x_k$ and $x = x_1, x_2, \dots, x_k$.

For illustration, suppose the number of categorical outcomes in Y is three, the logit transformation will have two functions. If it is assumed that $Y = 1$, then the functions follow.

$$g_1(x) = ln\frac{[P(Y=2|x]}{[P(Y=1|x]} = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1k}x_k \quad (2)$$

$$g_2(x) = ln\frac{[P(Y=3|x]}{[P(Y=1|x]} = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2k}x_k \quad (3)$$

One of the methods in estimating parameters for the multinomial logistic regression model is using the maximum likelihood method. The likelihood function used is:

$$L(\beta) = \prod_{i=1}^{n}\left[\pi_1(x_i)^{y_{1i}}\pi_2(x_i)^{y_{2i}}\pi_3(x_i)^{y_{3i}} \dots \pi_p(x_i)^{y_{pi}}\right] \quad (4)$$

The ln-likelihood function is obtained as follows [11]:

$$l(\beta) = \sum_{i=1}^{n} y_{1i}\, ln\, [\pi_1(x_i)] + y_{2i}\, ln\, [\pi_2(x_i)] + y_{3i}\, ln\, [\pi_3(x_i)] + \cdots + y_{pi}\, ln\, [\pi_p(x_i)] \quad (5)$$

So that the β can be estimated by maximizing the ln-likelihood function in equation (4). This can be done by the differentiation of the function to β then equal to zero or it can be done by Newton Raphson iteration.

The built model is then interpreted using the odds ratio value. The odds ratio and the model parameters have a relationship symbolized as (Ψ) with a value ($\Psi < 1$), meaning that the two variables have a negative relationship. It means that if the value of one variable increases, the other decreases. The two variables have a positive relationship if the value ($\Psi > 1$). Meanwhile, if $\Psi = 1$, it means that there is no relationship between two variables. Then this

relationship can be calculated using the following equation [12]:

$$\Psi = e^{\hat{\beta}} \tag{6}$$

observationsy variables in the model can be tested simultaneously using the G test statistic. This test has a null hypothesis. That is, there is no effect of the explanatory variable on the model ($H_0: \beta_1 = \cdots = \beta_k = 0$). If m is the index for the explanatory variables and $m = 1,2, \ldots, k$, then the alternative hypothesis is that at least one explanatory variable affects the model ($H_1: \beta_m \neq 0$) for $m = 1, \ldots, k$. $n_1, n_2, n_3, \ldots, n_k$ is the number of categorical observation. Systematically, this statistic of the G test can be calculated through the equation [13]:

$$G = -2\ ln\ \left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\left(\frac{n_3}{n}\right)^{n_3}\ldots\left(\frac{n_n}{n}\right)^{n_k}}{\prod_{i=1}^{n}\pi_1(x)^{y_{1i}}\pi_2(x)^{y_{2i}}\pi_3(x)^{y_{3i}}\ldots\pi_p(x)^{y_{pi}}}\right] \tag{7}$$

Additionally, the influence of explanatory variables in the model is partially tested using the Wald test. This test has a null hypothesis. That is, there is no effect of the explanatory variable on the model ($H_0: \beta_m = 0$). While the alternative hypothesis is that at least one explanatory variable affects the model ($H_1: \beta_m \neq 0$) for $m = 1, \ldots, k$. The standard error coefficient is symbolized by $\widehat{SE}(\hat{\beta}_m^2)$, where $\hat{\beta}_m$ is the estimated coefficient; systematically, this Wald test can be calculated using the equation [14]:

$$W^2 = \frac{\hat{\beta}_m^2}{\widehat{SE}(\hat{\beta}_m^2)} \tag{8}$$

2.3 Overdispersion in multinomial logistic regression

Overdispersion is a condition where the number of observations in the response variable is generally zero. The overdispersion can be detected by the following equation [15]:

$$\left(\phi = \frac{deviance\ value}{degree\ of\ freedom}\right) > 1 \tag{9}$$

We estimate parameters in the multinomial logistics regression ($\hat{\phi}$) based on Pearson fit statistics or ($\hat{\phi}_p$) can be determined as follows:

$$\hat{\phi}_p = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(y_{ij}-\hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}}}{N-n} \tag{10}$$

Where p is the total cells and for $y_i = \left(y_{i1}, \ldots, y_{ip}\right)^{\mathrm{T}}$ where $i = 1, .., n$ and the value of Nn is the total cells minus the degrees of freedom of the model by handling overdispersion due to n constraints [16].

2.4 Data Sources

One measure of the model's goodness-of-fit to the data is determined by the value of Akaike's information criterion. The goodness of this model is seen based on the value of Akaike's information criterion; if the value of Akaike's information criterion is smaller than the others, then the model is the best. The value of Akaike's information criterion can be calculated using the following equation [17], [18]:

$$AIC = 2k - 2\ ln\ ln\ (L) \tag{11}$$

Where:

k: the parameter of the model

L: the maximum value of the likelihood function used to estimate the model.

3 Result and Discussion

This study has focused on the multinomial regression method and its use in the overdispersion problem. We then used it to find the best model for the stress level problem experienced by students while studying online. The respondent variable was divided into four stress levels. Data were labeled as one if not experiencing stress (standard). Label two if the subject was experiencing mild stress, label three if the student was experiencing moderate stress, and label four for a student experiencing severe stress. The distribution of the stress level data is presented in Table 3 below.

Table 3. Class distribution of stress level data

Category	Class Size	Distribution Size
No stress (Normal)	115	40%
Experiencing Mild Stress	151	52%
Experiencing Moderate Stress	20	7%
Experiencing Severe Stress	2	1%

The stress level data was analyzed by multinomial regression with the seventeen factors presented in Table 1. The Stress Level Data model divides the data into training and test data. The training data used for

modeling is selected for a ratio of 80% of the data or 231 observations. Meanwhile, the testing data is used for the model evaluation. The ratio of 20% of the data or 57 observations is chosen.

Modeling using multinomial regression was carried out on these training data using a stepwise method to determine the explanatory variables that directly affect the stress level. The chosen model was based on the smallest AIC value of the model used. The results of the modeling using the stepwise method are shown in Table 4 below.

Table 4. The candidate model is the result of modeling with the stepwise method

Model	Explanatory Variables	AIC
1	$X_1 + X_2 + X_3 + X_9 + X_{10} + X_{11} + X_{12} + X_{13} + X_{14} + X_{15} + X_{16} + X_{17}$	343.58
2	$X_1 + X_3 + X_9 + X_{10} + X_{11} + X_{12} + X_{13} + X_{14} + X_{15} + X_{16} + X_{17}$	338.27
3	$X_1 + X_3 + X_9 + X_{10} + X_{11} + X_{12} + X_{13} + X_{14} + X_{16} + X_{17}$	335.67
4	$X_1 + X_3 + X_9 + X_{10} + X_{11} + X_{12} + X_{14} + X_{16} + X_{17}$	334.27
5	$X_1 + X_3 + X_{10} + X_{11} + X_{12} + X_{14} + X_{16} + X_{17}$	332.62

The best model obtained based on Table 4 is the model that contains the explanatory variables X_1, X_3, X_{10}, X_{11}, X_{12}, X_{14}, X_{16}, and X_{17}. This model has the smallest AIC value of 332.62 based on Table 4. The results of the modeling with multinomial stepwise logistic regression are presented in Table 5.

Table 5. Class distribution of stress level data

Model	Parameter	Estimation	Std. Error	P-Value
Multinomial Stepwise Logistic Regression	Intercept 1|2	0.2135	1.0437	0.2045
	Intercept 2|3	3.8305	1.0832	3.5362
	Intercept 3|4	6.1101	1.2753	4.7911
	X_1	-0.0759	0.0686	-1.106
	X_3	-0.8517	0.3725	-1.286
	X_{10}	0.3412	0.179	1.898
	X_{11}	0.2497	0.2350	1.063
	X_{12}	0.8748	0.4094	2.136
	X_{14}	0.6871	0.3022	2.273
	X_{16}	0.9344	0.3130	2.985
	X_{17}	1.1658	0.3955	2.948

The chi-squared value using the G test was obtained at 175.455203, and the result was significant at the 5% level. The model fits the data where at least one explanatory variable affects the response variable. Furthermore, the Wald test is performed and this test is based on the p-value. This test aims to determine the effect of each explanatory variable on the response variable. Based on the best model, the variables X_1, X_3, X_{10}, X_{11}, X_{14}, X_{16}, and X_{17} significantly affect the response variable Y in terms of building the best model.

The multinomial stepwise logistic regression model was then evaluated using classification training and test data. The multinomial stepwise logistic regression model was built using the training data and then tested; it was found to have a precision of 77.19% for the classification results. So this model is feasible to use.

Further testing of the best multinomial logistic regression model shows that the chi-square value with the deviation-free degree has more than one value. It indicates that there is an overdispersion. Therefore, the overdispersion treatment will be carried out next. The result of multinomial logistic regression modeling with overdispersion handling can be seen in Table 6 below.

Table 6. Modeling result of multinomial logistic regression modeling with overdispersion handling

Model	Parameter	Estimation	Std. Error	P-Value
Experiencing Mild Stress	Intercept	2.63830	1.49180	0.077430
	X_1	-0.15372	0.07993	0.054879
	X_3	-2.30360	0.65116	0.000432
	X_{10}	0.14452	0.24503	0.555518
	X_{11}	0.19364	0.32305	0.549110
	X_{12}	1.06225	0.50911	0.037316
	X_{14}	0.9076	0.3865	0.0191

		0	3	64
	X_{16}	1.74668	0.39905	1.4e-05
	X_{17}	3.07448	1.09615	0.005182
Experiencing Moderate Stress	Intercept	-2.40396	2.34923	0.30654
	X_1	0.09425	0.12262	0.44239
	X_3	-2.14377	0.89697	0.01712
	X_{10}	0.57577	0.36685	0.11701
	X_{11}	0.90824	0.48788	0.06310
	X_{12}	-0.02685	0.76843	0.97213
	X_{14}	1.12216	0.69926	0.10902
	X_{16}	0.92648	0.69246	0.18136
	X_{17}	3.99013	1.21317	0.00106
Experiencing Severe Stress	Intercept	-859.42	176267.83	0.996
	X_1	-20.21	1847.57	0.991
	X_3	140.77	80984.80	0.999
	X_{10}	90.67	10197.84	0.993
	X_{11}	110.07	15764.56	0.994
	X_{12}	68.60	15446.45	0.996
	X_{14}	109.44	13638.97	0.994
	X_{16}	31.57	11683.97	0.998
	X_{17}	-7.91	17824.23	1.000

Based on the results of estimating the parameters of the multinomial logistic regression model, with the handling of overdispersion as above, the multinomial logistic regression model based on the logit function can be written as follows:

$$g_{mildstress}(x) = 2.63830 - 0.15372\ X_1 - 2.30360\ X_3 + 0.14452\ X_{10} + 0.19364\ X_{11} + 1.06225\ X_{12} + 0.90760\ X_{14} + 1.74668\ X_{16} + 3.07448\ X_{17}$$

$$g_{moderatestress}(x) = -2.40396 + 0.09425\ X_1 - 2.14377\ X_3 + 0.57577\ X_{10} + 0.90824\ X_{11} - 0.02685\ X_{12} + 1.12216\ X_{14} + 0.92648\ X_{16} + 3.99013\ X_{17}$$

$$g_{severestress}(x) = -859.42 - 20.21\ X_1 + 140.77\ X_3 + 90.67\ X_{10} + 110.07\ X_{11} + 68.60\ X_{12} + 109.44\ X_{14} + 31.57\ X_{16} - 7.91\ X_{17}$$

Based on the results of the modeling in Table 6, which shows the students who experience mild stress and the explanatory variables that are significant. The odds ratio value can be interpreted as follows. If the stress level experienced by students is mild, then the average chance of students using zoom is 0.15372 times less than that of students that are not stressed. Suppose that the stress level experienced by students is low. In that case, the probability of a small device type is 2.30360 times less than for students who do not experience stress. Suppose the stress level experienced by the student is mild. In that case, the probability of experiencing eye fatigue is 1.06225 times higher. Then the probability of experiencing a tingling sensation is 0.90760 times greater. The chance of experiencing headaches is 1.74668 times more significant, and the probability of experiencing finger pain is 3,07448 times greater than students who do not experience stress.

Meanwhile, suppose that the stress level experienced by the students is moderate. In that case, the chances of a small device type are 2,14377 less than students who do not experience stress. Suppose that the stress level experienced by students is moderate. In that case, the chance of experiencing finger pain is 3.99013 higher than students who do not experience stress.

For the evaluation model, the multinomial logistic regression model with overdispersion handling is the best based on the AIC compared without accommodating the overdispersion case. It can be seen in Table 7 below.

Table 7. The evaluation of the modeling result of multinomial logistic regression

Model	AIC
Multinomial logistic regression	332.62
Multinomial logistic regression with handling of overdispersion	167.38

4 Conclusion

On the basis of the analysis that has been carried out, we believe that data modeled using overdispersion treatment gives a better performance. However, both

models have explanatory variables that significantly affect the same response variable: average usage, device type, audio type, usage limit, eye condition, tingling, head condition, and finger condition. In the future, research can include a random effect to add the model to evaluate unobserved variables.

Acknowledgment:
No funding was received for the conduct of this study.

Conflict of Interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.

References:

[1] H. Ij, 'Statistics versus machine learning," *Nat Methods*, vol. 15, no. 4, p. 233, 2018.

[2] T. B. Ambo, J. Ma and C. Fu, "Investigating influence factors of traffic violation using multinomial logit method", *Int J Inj Contr Saf Promot*, vol. 28, no. 1, pp. 78–85, 2020.

[3] M. R. Adha, S. Nurrohmah, and S. Abdullah, 'Multinomial Logistic Regression and Spline Regression for Credit Risk Modeling', in *Journal of Physics: Conference Series*, 2018, vol. 1108, no. 1, p. 012019.

[4] V. Landsman, D. Landsman, C. S. Li, and H. Bang, 'Overdispersion models for correlated multinomial data: Applications to blinding assessment", *Stat Med*, vol. 38, no. 25, pp. 4963–4976, 2019.

[5] E. Castilla, N. Martn and L. Pardo, "Minimum phi-divergence estimators for multinomial logistic regression with complex sample design', *AStA Advances in Statistical Analysis*, vol. 102, no. 3, pp. 381–411, 2018.

[6] G. Fauville, M. Luo, A. C. M. Queiroz, J. N. Bailenson, and J. Hancock, 'Zoom exhaustion & fatigue scale," *Computers in Human Behavior Reports*, vol. 4, p. 100119, 2021.

[7] E. Peper, V. Wilson, M. Martin, E. Rosegard and R. Harvey, "Avoid Zoom fatigue, be present and learn', *NeuroRegulation*, vol. 8, no. 1, p. 47, 2021.

[8] V. I. Manea, T. Macavei, and C. Pribeanu, 'Stress, frustration, boredom, and fatigue in online engineering education during the pandemic', *International Joural of User-System Interaction*, Vol. 13, no. 4, pp. 199–214, 2020.

[9] A. Bonanomi, F. Facchin, S. Barello, and D. Villani, 'Prevalence and health correlates of onine fatigue: A cross-sectional study on the Italian academic community during the COVID-19 pandemic," *PLoS One*, vol. 16, no. 10, p. e0255181, 2021.

[10] Y. Bayar, H. F. Sezgin, Ö. F. Öztürk, and M. Ü. Şaşmaz, "Financial literacy and financial risk tolerance of individual investors: Multinomial logistic regression approach, *Sage Open*, vol. 10, no. 3, p. 2158244020945717, 2020.

[11] H. B. Khudair, K. G. Khalid and K. R. Jbbar, "Condition Prediction Models of Deteriorated TrunkSewer Using Multinomial Logistic Regression and Artificial Neural Network," *Int. J. Civ. Eng. Technol.* vol. 10, pp. 93–104, 2019.

[12] S. Buya, P. Tongkumchum and B. Owusu, "Modelling of land use change in Thailand using binary logistic regression and multinomial logistic regression", *Arabian Journal of Geosciences*, vol. 13, p. 437, Jun. 2020, doi: 10.1007/s12517-020-05451-2.

[13] J. Lee, S. Yasmin, N. Eluru, M. Abdel-Aty, and Q. Cai, "Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects", *Accid Anal Prev*, vol. 111, pp. 12–22, 2018.

[14] A. Abdillah, A. Sutisna, I. Tarjiah, D. Fitria, and T. Widiyarto, 'Application of Multinomial Logistic Regression to Analyze Learning Disabilities in statistics courses," in *Journal of Physics: Conference Series*, 2020, vol. 1490, no. 1, p. 012012.

[15] N. Corsini and C. Viroli, "Dealing with overdispersion in multivariate count data", *Comput Stat Data Anal*, vol. 170, p. 107447, Feb. 2022, doi: 10.1016/j.csda.2022.107447.

[16] F. Afroz, M. Parry and D. Fletcher, "Estimating overdispersion in sparse multinomial data', *Biometrics*, vol. 76, no. 3, pp. 834–842, 2020.

[17] J. E. Cavanaugh and A. A. Neath, 'The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdiscip Rev Comput Stat*, vol. 11, no. 3, p. e1460, 2019.

[18] H. De and G. Acquah, 'Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in the selection of an asymmetric price relationship', *J Dev Agric Econ*, vol. 2, pp. 1–6, Feb. 201

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
A'yunin Sofro is the coordinator
Khusnia Nurul Khikmah has written an original draft
Danang Ariyanto was responsible for the review and editing.
Yusuf Fuad was responsible for the review.
Budi Rahadjeng was responsible for the review.
Yuliani Puji Astuti was responsible for the review