Mathematics and Statistics 11(2): 379-384, 2023 http://www.hrpub.org
DOI: 10.13189/ms.2023.110217

Binary Response on Logistics Regression Model and Its Simulation

Budi Pratikno[1,*], Fifthany Marchelina Napitupulu[1,2], Jajang[1], Agustini Tripena Br. Sb[1], Mashuri[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science, Jenderal Soedirman University, Purwokerto, Indonesia
[2]Jl. Dr. Soeparno 61 Karangwangkal, Purwokerto, 53122, Central of Java, Indonesia

Received December 18, 2022; Revised January 20, 2023; Accepted March 21, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] Budi Pratikno, Napitupulu F.M., Jajang, Tripena A, Mashuri , "Binary Response on Logistics Regression Model and Its Simulation," Mathematics and Statistics, Vol. 11, No. 2, pp. 379 - 384, 2023. DOI: 10.13189/ms.2023.110217.*

(b): *Budi Pratikno, Napitupulu F.M., Jajang, Tripena A, Mashuri (2023). Binary Response on Logistics Regression Model and Its Simulation. Mathematics and Statistics, 11(2), 379 - 384. DOI: 10.13189/ms.2023.110217.*

Abstract The research determined the binary response model on logistics regression (LR) and its application. Firstly, we select some eligible factors (predictors, $X_i,\ i=1,2,3,4$) that are involved in the model, namely age (X_1), sex (X_2), treatment (X_3), and nutrition (X_4), with the response (Y) being the case of tuberculosis (TB). Using the stepwise selection model and odd ratio (*OR*) interpretation, we have three suspected significant predictors (X_1, X_3, and X_4), but we choose two (only) of the significant predictors, which are X_3 and X_4. Therefore, the logistics regression model is written as $p(x)=\frac{e^{g(x)}}{1+e^{g(x)}}=\frac{e^{-2.2+1.6X_3+1.1X_4}}{1+e^{-2.2+1.6X_3+1.1X_4}}$. To test the goodness of fit of the model, we used deviance test (*p-value* ≈ 0.08). Due to this *p-value*, we then used the level of significance which is 0.08 (nearly close to 0.05) for obtaining the significant model. For more detailed interpretation, we here noted that the OR of the age (X_1), one of the three suspected significant predictors (X_1, X_3, and X_4), is close to be one (≈ 1), so it is an independent predictor (not significant). So, we concluded that the significant predictors are only treatment (X_3) and nutrition (X_4). Thus, the linear of the logistics regression model is then given as a $\ln\left(\frac{p(x)}{[1-p(x)]}\right)=-2.2+1.6X_3+1.1X_4$. So, we noted that TB is only dependent on clinical treatment and providing nutrition.

Keywords Binary Logistics Regression, Deviance, Odds Ratio, Medical Case

1. Introduction

In the context of inferential statistics, population inference is usually drawn from sample. How the population inference could be improved? There are some statistical techniques to improve the population inference such as the use of non-sample prior information (NSPI) on the maximum power and minimum size, and decision theory on the binary response model. In terms of the NSPI context, many authors studied improving inference population, such as Pratikno [3], Khan and Pratikno [14], Khan [15], Khan and Saleh [16], Khan and Hoque [17], Saleh [2], Yunus [12], and Yunus and Khan [13]. In the context of the decision theory, many authors such as Melsa and Cohn [9], Yan and Blum [19], and Trees [8], studied the binary choices (right or false conclusion) model. Here, we noted that Yan and Blum [14] and Trees [8] already studied many applications of the Neyman-Pearson (NP) in medical cases of heart attack, and Melsa and Cohn [9] have discussed the maximum likelihood (ML) on a binary case that follows *Gaussian* or Normal (Z) distribution. Furthermore, Montgomery [6] and Chase and Jacobs [11] also studied the estimation model of data series.

In the context of the binary responses model (*Y*), the multiple regression models (MRM) are not a suitable model, so the logistics regression (LR) model is then used. Here, the binary (categories) responses, namely 0 (failure) and 1 (success) must be modeled by LR [7]. By definition, we noted that the binary logistics regression is a method that can be used to find the relationship between one dependent variable (binary responses, *Y*) with one or more independent variables (*X*) [1] [7] [10]. For more detailed interpretation, we also used the odds ratio (OR) to interpret the relationship between binary responses and predictors in this model. Moreover, we used the maximum likelihood estimation (MLE) method, the test statistics ***G*** (the likelihood ratio test or LRT) and Wald (*W*) test, to test the coefficient regression parameter and the goodness of fit of the model.

In this paper, Section 1 presents the introduction. The research methodology is given in Section 2. The previous research and theory are given in Section 3. The result is obtained in Section 4, and Section 5 describes the conclusion of the research.

2. Methods

Step 1. We studied the binary responses, Binomial distribution and logistics regression model.

Step 2. We choose the eligible method for the binary responses model, namely the logistics regression model and odd ratio (OR).

Step 3. For simulation, we used data in Table 1.

Table 1. Name of Variables

Notation	Name of Variable	Remark
Y	Tuberculosis (TB)	0 = fail 1 = Success
X_1	Age	
X_2	Sex	0 = Male 1 = female
X_3	Treatment	0 = Not 1 = Yes
X_4	Nutrition	0 = lack nutrition 1= full nutrition

Due to the previous research, the binary response is often applied in the topic of the clinical trial, so we used data in Table 1 as a simulation in which *Y* has a binary response.

To test the validity data, the correlation coefficient (*r*) is used, that is,

$$r = \frac{\sum_{i=}^{n} X_i Y_i - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\right)\left(\sum_{i=1}^{n} Y_i^2 - \left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)\right)}}.$$

Using central limit theorem, for large *n* $(n \geq 30)$, the negative correlation *r* should be -1 < *r* < -0.3 and the positive correlation *r* should be 0.3 < *r* < 1. Hence, we considered the validity testing data when *r* < - 0.3 and *r* > 0.3

Furthermore, the reliability data is tested using Cronbach's Alpha. It is computed using the score from each scale item and correlating them with the total score for each observation.

{\displaystyle \alpha =({k \over k-1})\left(1-{\sum _{i=1}^{k}\sigma _{y}^{2} \over \sigma _{x}^{2}}\right)}

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{i=1}^{n} \sigma_y^2}{\sigma_x^2}\right)$$

When the Cronbach's Alpha is greater than 0.6, the reliability data is significant.

3. Previous Research and Theory

Many authors studied binary responses on the logistics regression (LR) model such as Hosmer, et al. [7], Nirwana and Sukarna [18], Rahmadeni and Safitri, [10], and Agresti [1]. The simple model of the LR is written as

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}, \tag{1}$$

with $e = 2.718...$, $p(x_i)$ is probability of success, and β_i coefficient regression parameter. To learn more details about the logistics regression model, we learned the multiple regression model as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e_i$$

with Y is response, $i = 1, \cdots, n$, $\beta_0, \beta_1, \cdots, \beta_k$ are the coefficient regression parameters, $X_1, \cdots, X_k$ are predictors, and e_i is error. Furthermore, Collect [4] presented the logistics regression model as

$$p(x)=\frac{e^{g(x)}}{1+e^{g(x)}}=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}. \quad (2)$$

with $e=2.718...$, and $p(x)$ is probability of success, the linear equation of equation (2) is then given as a $\ln\left(\frac{p(x)}{\left[1-p(x)\right]}\right)$ as below.

$$p(x)\left[1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}\right]=e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$$

$$\Leftrightarrow p(x)+p(x)e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}=e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$$

$$\Leftrightarrow p(x)=e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}-p(x)e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$$

$$\Leftrightarrow p(x)=\left[1-p(x)\right]e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$$

$$\Leftrightarrow \frac{p(x)}{\left[1-p(x)\right]}=e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}, \text{ taking ln}$$

$$\Leftrightarrow \ln\left(\frac{p(x)}{\left[1-p(x)\right]}\right)=\ln\left(e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}\right)$$

$$\therefore \ln\left(\frac{p(x)}{\left[1-p(x)\right]}\right)=\beta_0+\beta_1X_1+\cdots+\beta_kX_k.$$

Referring to Kleinbaum and Klein [5], and using minimize of the maximum likelihood estimation (MLE) of the likelihood function $L(\beta)$, we then got an estimator of the coefficient regression parameters, $\widehat{\beta}$.

To test the coefficient regression parameters β , we used two common methods, namely, the *ratio likelihood test (G)* in testing $H_0:\beta_1=\cdots=\beta_k=0$. If $G>\chi^2_{(\alpha;k)}$, H_o is rejected, where

$$G=-2\ln\left(\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}_i^{\,y_i}\left(1-\widehat{\pi}_i\right)^{(1-y_i)}}\right)$$

$n=n_1+n_0$, n is number of trials, n_1 is number of trials on *Y=1*, n_0 is number of trials on *Y=0*, $\chi^2_{(\alpha;k)}$ is chi-square distribution, and $\widehat{\pi}_i$ is probability of success [7], and the *Wald test* is used for partial testing, in testing $H_0:\beta_k=0$. H_o is rejected when

$$W=\left(\frac{\widehat{\beta}}{SE\left(\widehat{\beta}\right)}\right)^2>\chi^2_{(\alpha;1)}$$

Furthermore, we test the goodness of fit of the model using *deviance* testing. Here, H_o is rejected when $D>\chi^2_{(\alpha;n-k)}$, where H_0 is goodness of fit and

$$D=-2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right)+\ln\left(1-\widehat{\pi}_i\right)\right].$$

Following Hosmer, et al. [7], we presented the odd ratio (OR) as

$$OR=\frac{\left(\frac{\pi(1)}{1-\pi(1)}\right)}{\left(\frac{\pi(0)}{1-\pi(0)}\right)}=\left(\frac{\frac{\frac{e^{\beta_0+\beta_1x_1}}{1+e^{\beta_0+\beta_1x_1}}}{1-\frac{e^{\beta_0+\beta_1x_1}}{1+e^{\beta_0+\beta_1x_1}}}}{\frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{1-\frac{e^{\beta_0}}{1+e^{\beta_0}}}}\right)=\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}=e^{\beta_1} \quad (3)$$

The equation (3) or odd ratio (OR) is used as interpretation of the probability model of the $p(x)$, that are: (1) $OR<1$, it means that the probability success (Y=1) is less than the probability failure (Y=0), (2) $OR=1$, the independent case of (Y=1) and (Y=0) occurs, and (3) $OR>1$, the probability success (Y=1) is greater than the probability of failure (Y=0) [1]. For example, if the $\beta_0=-5.3$ and $\beta_1=0.1$, then the OR is given as $\exp(\beta_1)=\exp(0.1)=1.1$. It means that the OR is greater than one, so the probability of success (Y=1) is significant.

4. Result

To illustrate the power, we let X_i as a random variables fallow Bernoulli distribution with parameter θ, $n=20$, and $Y=\sum_{j=1}^{n=20}X_j$, $Y\sim \text{Bin}(n,\theta)$. The power of this distribution, in testing $H_0:\theta=0.5$ versus $H_1:\theta\neq 0.5$, on rejection area $\{(x_1,...,x_{20}):\text{Y}\leq 6\}$, is given as

$$\pi(\theta) = P(\text{reject } H_0 \mid \text{under } H_1) = \sum_{y=0}^{6} \binom{20}{y} \theta^y (1-\theta)^{20-y} \Big|_{\theta \neq 0.5}.$$

Following the previous theory, we must choose the maximum power and minimum size to obtain the right conclusion**,** where the size is given as

$$\alpha(\theta) = P(\text{reject } H_0 \mid \text{under } H_0) = \sum_{y=0}^{6} \binom{20}{y} \theta^y (1-\theta)^{20-y} \Big|_{\theta = 0.5}.$$

In the context of the binary responses in Table 1, the power of the Binomial case is not suitable, so the power could not be used for this data. We then determined the logistics regression model as a significant method for the binary responses problem for Table 1. The estimate parameters model is then given in Table 2.

Table 2. The estimate parameters model using *R*

Predictors	Parameters	Estimate
X_1	β_1	0.03
X_2	β_2	0.55
X_3	β_3	1.66
X_4	β_4	1.10
constant	β_0	-2.73

The *R-code* of the output in Table 2 is presented as reg<-glm(y ~ x1+x2+x3+x4, family=binomial), summary(reg).

Furthermore, we test the hypothesis coefficient regression parameters, $H_0 : \beta_1 = \cdots = \beta_k = 0$, using

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \widehat{\pi}_i^{\,y_i} \left(1-\widehat{\pi}_i\right)^{1-y_1}}\right] \quad (4)$$

Using equation (4) and *R-code,* the *G* is 31.7, the degree of freedom is 4, and the *p-value* is 0.00. Here, the *Chi-square* $\chi^2_{0.05;4}$ = 9.5*,* so we reject H_0. The test statistics of the *G* is presented in Table 3.

Table 3. The *G* Test and Its *p-value*

Software	G Test	Degree of freedom	p-value
R	31.7	4	0.0

The partial testing (*Wald-test*) is then used to find the significant variables involved in the model (see Table 4).

Table 4. The Wald Test and Its *p-values*

Predictors	*Wald*	*p-value*
X_1	2.2	0.03
X_2	1.3	0.20
X_3	4.0	0.00
X_4	2.8	0.01

From Table 4, (*R-code* output), the X_3 and X_4 are *significant*, with the *p-value* = 0.00 and *p-value*= *0.01*, respectively. Thus, X_3 and X_4 are significant, and the estimate parameters model of the X_3 and X_4 are given in Table 5.

Table 5. The estimate parameters model

Predictors	Parameters	Estimate
X_3	β_3	1.6
X_4	β_4	1.1
constant	β_0	-2.2

From Table 5, we noted that the best model of the LR and its linear model are

$$p(x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{-2.2+1.6X_3+1.1X_4}}{1+e^{-2.2+1.6X_3+1.1X_4}}, \text{ and}$$

$$\ln\left(\frac{p(x)}{[1-p(x)]}\right) = -2.2 + 1.6X_3 + 1.1X_4, \text{ respectively.}$$

Moreover, we then test the goodness of fit of the model. The deviance test (*p-value*= *0.08*) is used. Due to the *p-value* of the deviance test 0.08, we, therefore, choose the level of significance 0.08 (nearly 0.05), so the model is significant. Similarly, H_0 is not rejected when $\widehat{C} = 7.8 < \chi^2_{(\alpha;g-2)} = 15.5$ (Hosmer and Lemeshow test), where

$$\widehat{C} = \sum_{i=1}^{g} \frac{\left(o_{1i} - \widehat{e}_{1i}\right)^2}{\widehat{e}_{1i}} + \frac{\left(o_{0i} - \widehat{e}_{0i}\right)^2}{\widehat{e}_{0i}}$$

with *g* = *10* is number of group, O_{1i} is total observation on *Y=1,* 0_{0i} is total observation on *Y=0* , $\widehat{e}_{1i}$ is expected value on *Y=1*, $\widehat{e}_{0i}$ is expected value on *Y=0*. From both testing, we then concluded that the model is *goodness of fit*. The output of the deviance and Hosmer and Lemeshow test is obtained in Table 6.

Table 6. The Output deviance, and Hosmer and Lemeshow

Testing	*software*		*Chi-square*
	R	***p-value***	
Deviance	160	0.08	164.2
Hosmer and Lemeshow	7.8	0.45	15.5

We see from Table 6 that the *p-value* of the deviance is smaller than the *p-value* of the Hosmer-Lameshow, so we prefer using the deviance testing as an indicator testing of *the goodness of fit*. Here, we used $\alpha = 0.08$ (close to 0.05), and the *p-value* of the Hosmer and Lemeshow is 0.452 (too high, greater than 0.08).

For a more detailed interpretation of this research's result, we presented the odd ratio (OR) on three suspected significant variables in Table 7.

Table 7. The OR of the suspected significant variables

Variable	*Odds ratio*
X_1	1.0
X_3	4.9
X_4	2.9

From Table 7, it is clear that age is independent (not significant) to the recovery of the TB, the OR is 1.03 close to one, but both treatment and nutrition (Their OR > 1) are significant for recovery patient of the TB.

From several tests and selected mechanisms above, we then presented the final eligible model as $p(x) = \frac{e^{-2.2+1.6X_3+1.1X_4}}{1+e^{-2.2+1.6X_3+1.1X_4}}$. Here, we noted that the relationship between the model and the outcome of the simulation is TB only depended on clinical treatment and providing nutrition, in the scope of area data intake.

Noted that the output in Table 1 till Table 6 is produced using *R* and/or *Minitab*. The *R-code* and *Minitab command* are given as follows

(1) getwd()
[1] 'C'
>setwd ('C')
Df<read.table("Ranalysis.txt", header=TRUE)
view (df)
glm.fit<-glm(Y ~ $\sum X_i$, data=df, family=binomial)
summary(glm.fit)
>library(ResourceSelection)
>hoslem.test(reg$Y, fitted(Y), g=10)
>exp(coef(Y))

(2) Click *Stat*, *Regression*, *Binary Logistic Regression*, and *Fit Binary Logistic Model.*

5. Conclusions

The research studied the binary responses model. We used logistics regression method to select an eligible factor (predictor) involved in the model. Using the stepwise method and odd ratio (OR), we got the significant predictors X_3 and X_4, so the model is written as

$$p(x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{-2.2+1.6X_3+1.1X_4}}{1+e^{-2.2+1.6X_3+1.1X_4}}.$$

To test the goodness of fit of the model, we used deviance test (*p-value* = 0.08). Due to this *p-value*, we then used the level of significance 0.08 (nearly close to 0.05) for obtaining the significant model. For more detailed interpretation, we used the OR for three suspected significant predictors (X_1, X_3, and X_4), with the OR of the age (X_1) close to be one (≈ 1), so it is independent predictor (not significant). Finally, we concluded that the significant predictors are only treatment (X_3) and nutrition (X_4), so the eligible model is $p(x) = \frac{e^{-2.2+1.6X_3+1.1X_4}}{1+e^{-2.2+1.6X_3+1.1X_4}}$. Finally, we noted that TB is only depended on clinical treatment and providing nutrition, but this conclusion is only used in the scope of the area's data intake, one of the hospital in Banyumas.

Acknowledgements

I would like to thank the LPPM UNSOED for providing me granting of research.

REFERENCES

[1] A. Agresti. Categorical Data Analysis, Second Edition. John Wiley & Sons, New York, 2002.

[2] A. K. Md. E. Saleh. Theory of preliminary test and Stein-type estimation with applications. John Wiley and Sons, Inc., New Jersey, 2006.

[3] B. Pratikno. Test of Hypothesis for Linear Models with Non-Sample Prior Information. Unpublished PhD Thesis, University of Southern Queensland, Australia, 2012.

[4] D. Collet. Modelling Binarry Data. Second Edition. England: Taylor & Francis, 2003.

[5] D.G. Kleinbaum and M. Klein. Logistic Regression: A Self Learning Text (3rd ed). Springer, New York, 2010.

[6] D. Montgomery. Forecasting and Time Series Analysis. Jhon Wiley and Sons. Inc., New Jersey, 2008.

[7] D.W. Hosmer, S. Lemeshow, and R. X. Sturdivant. Applied Logistic Regression. Third Edition. John Wiley and Sons Inc., Canada, 2013.

[8] H.L.V. Trees. Detection, Estimation, and Modulation, Part 1. A Wiley-Inter science, Canada, 2001.

[9] J. L. Melsa and D. L. Cohn. Decision and Estimation Theory. McGraw-Hill Kogakusha, Toky, 1978.

[10] Rahmadeni and E. Safitri. Pemodelan Pasien Kanker Payudara Menggunakan Regresi Logistik Biner (Studi Kasus: Pasien Kanker Payudara di Rumah Sakit Umum Daerah Arifin Ahmad Pekanbaru). Pekanbaru. Jurnal Sains, Teknologi dan Industri 13(2): 168-173, 2016.

[11] R. Chase and R. Jacobs. Operation and Supply Chain Management. Global Case Edition, Mc Graw Hill, New York, 2014.

[12] R. M Yunus. Increasing power of M-test through pre-testing. Unpublished PhD Thesis, University of Southern Queensland, Australia, 2010.

[13] R. M. Yunus and S. Khan. Test for intercept after pre-testing on slope a robust method. In: 9th Islamic Countries Conference on Statistical Sciences (ICCS-IX): Statistics in the Contemporary World - Theories, Methods and Applications, 2007.

[14] S. Khan and B. Pratikno. Testing Base Load with Non-Sample Prior Information on Process Load. Statistical Papers, 54 (3), 605-617, 2013.

[15] S. Khan. Estimation of the Parameters of two Parallel Regression Lines Under Uncertain Prior Information. Biometrical Journal, 44, 73-90, 2003.

[16] S. Khan and A. K. Md. E. Saleh. Preliminary test estimators of the mean based on p-samples from multivariate Student-t populations. Bulletin of the International Statistical Institute. 50th Session of ISI, Beijing, 599-600, 1995.

[17] S. Khan and Z. Hoque. Preliminary test estimators for the multivariate normal mean based on the modified W, LR and LM tests. Journal of Statistical Research, Vol 37, 43-55, 2003.

[18] Sukarna and S.R.A. Nirwana. Aplikasi Regresi Logistik Multinomial dalam Menentukan Faktor-faktor yang Mempengaruhi Pemilihan Program Studi di Jurusan Matematika FMIPA UNM. Journal of Mathematics and statistics, 1(1): 65-72, 2015.

[19] Q. Yan and R.S. Blum. Distributed Signal Detection Under the Neyman-Pearson Criterion. Journal of IEEE Transactions on Information Theory 47(4), 2001.