Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2023;12(4):1180-4

Effects of multicollinearity on type I error rate and test power of binary logistic regression model: A simulation study

Yeliz Kasko Arici[1], Mustafa Muhip Ozkan[2], Zahide Kocabas[2]

[1]*Ordu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Ordu, Türkiye*
[2]*Ankara University, Department of Biometry and Genetics, Ankara, Türkiye*

Received 16 August 2023; Accepted 14 September 2023
Available online 09.10.2023 with doi: 10.5455/medscience.2023.08.146

Abstract

In this study, the effect of multicollinearity, which is defined as high correlation, on the type I error rate and test power of the binary logistic regression models were studied. To do this, one dependent variable that consists of 1 and 2 and four continuous independent variables that were randomly drawn from the standardized normal distribution were taken into consideration in the constructed binary logistic regression model. To calculate the type I error rates and test power, the simulation study was performed 100.000 times. The simulation study repeated for the sample sizes of 10, 20, 30 and 40 the various degrees of correlations, namely 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%. In order to calculate test power, differences (δ) were created between population means by adding to the mean of the second population in standard deviation units as 0.5, 1.0, 1.5 and 2.0, respectively. The simulation runs exhibited that the increasing degree of multicollinearity among independent variables had no influence on type I error rates, provided that the sample size should not be smaller than 30. The power of the binary logistic regression was least affected by the increasing degrees of multicollinearity when the sample size is 10 and there is a 0.5δ difference between population means. The fact that there was a marked decline in test power with rising multicollinearity for all sample sizes clarified that the binary logistic regression was most powerful if there is no strong multicollinearity among independent variables when there is a 1.0δ difference between population means. The negative impact of the rising degrees of multicollinearity on the test power can be avoided if the number of observations is sufficiently large and if the populations were satisfactorily separated from each other.

Keywords: Logistic regression, binary response variable, multicollinearity, type I error rate, test power

Introduction

Regression analysis investigates the relationship between a dependent variable and one and/or more than one independent variable. It also provides information on variation in dependent variables explained by independent variables. To obtain reliable results from the standard regression analysis, the assumptions its must be accomplished in the data. The most important assumptions are that the dependent variable should be continuous and have a normal distribution and there should not be collinearity among independent variables [1,2]. However, in most studies the dependent (outcome) variable may be binary, consisting of two components or elements, or dichotomous, dividing the data into two sharply distinguished parts or classifications. Under this circumstance, standard regression analysis cannot be used to investigate the relationship between dependent and independent variables [3,4].

When the dependent variable is not continuous, logistic regression is a reliable method to investigate the relationship. The logistic regression analysis distinguishes from the standard regression analysis in terms of the type of dependent variable. While logistic regression requires a binary or dichotomous dependent variable, standard regression analysis needs a continuous dependent

CITATION
Kasko Arici Y, Ozkan M, Kocabas Z. Effects of multicollinearity on type I error rate and test power of binary logistic regression model: A simulation study. Med Science. 2023;12(4):1180-4.

Corresponding Author: Yeliz Kasko Arici, Ordu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Ordu, Türkiye
Email: yelizkasko@yahoo.com

DOI: 10.5455/medscience.2023.08.146

variable. This difference between them impacts both choices of parameters and assumptions [3,4].

Logistic regression analysis is commonly used due to the fact that the results of it can be interpreted without difficulty and it does not need any assumption related to independent variables. The method was first recommended to use in biology by Berkson [5-7] and has become more popular with the use of discriminant function [8]. The logistic regression method which was applied in various fields [9] was recommended as an alternative to discriminant analysis when independent variables did not normally distribute [10] and to probit analysis [11]. It was reported that the estimation of parameters in the logistic model was not sufficiently accurate when there is a noticeable multicollinearity among independent variables, which could cause misinterpretation of the odds ratio. Moreover, depending on a simulation study, the reported there was indecisiveness in logistic regression under multicollinearity, which causes to decide that all estimates of parameters were statistically insignificant even if all independent variables were included in the model. However, the results of goodness-of-fit test were in contrast with the estimation of parameters [12].

In another study conducted, the effect of collinearity on logistic regression was studied. It was taken X_1 and X_2 which were highly correlated with each other, and X_3 which was almost uncorrelated with the others from a uniform distribution. The results of the study showed that the collinearity between X_1 and X_2 resulted in an increment in standard error when only X_1 and X_2 were included in the logistic model. The acceptable results were attained when the model was built by only using X_3. Moreover, when all the variables were used in the model, the results were a mixture of the above-mentioned models and satisfactory, which emphasized that multicollinearity affected the reliability of the results [3]. It was also shown that if one independent variable was a perfect linear function of another in the model, standard errors became infinite and the solution to the model was indeterminate [13]. On the other hand, multicollinearity may also be a problem in logistic regression when one independent is a close-perfect linear function of another independent. The logit coefficients' standard errors are inflated as the independent variables' correlation with one another increases. Only their reliability is affected by multicollinearity, not the estimates of the coefficients [14].

In literature, there are plenty of studies stating the effect of multicollinearity on logistic regression analysis depending on the real observations even though logistic regression is assumed not to require any certain assumption [4,15]. If the calculated coefficients and standard errors are higher than expected, the results of hypothesis control will change. The effects of multicollinearity on the binary logistic regression model's standard errors and coefficients have also been the subject of numerous studies [13-15]. However, it appears that very few studies have looked at what happens to type I error rates when a binary logistic regression model contains multicollinearity. The purpose of this simulation study is to investigate the impact of multicollinearity on the type I error and test power in binary logistic regression analysis. In the study, type I error and power of test were observed under the changing degree of multicollinearity among four independent variables, different sample sizes and various differences as standard deviation between populations and were compared with each other.

Material and Methods

Logistic regression analysis enables researchers to determine the relationship between binary (dichotomous) dependent a variable and a number of independent variables that contain both discrete and continuous.

The logit of the multiple logistic regression models including p independent variables is written as in Eq. (1) [3]:

$$g(x)=b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p \tag{1}$$

In this case the logistic regression model is defined by Eq. (2) [3]:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{2}$$

Where p(x)=P(Y=1|X), which is also known as Odds ratio. The unknown regression coefficients in the logistic regression model are estimated as in Eq. (3) [3]:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \tag{3}$$

After the regression coefficients are estimated, the observed values are compared with the predicted values using the likelihood function in Eq. (4) [3]:

$$D = -2\ln\left(\frac{\text{likelihood of the the fitted model}}{\text{likelihood of the saturated model}}\right) \tag{4}$$

To investigate whether or not it is worthwhile to include an independent variable in the model is examined by comparing the D values with and without the variable. The change in the D due to the addition of the independent variable in the model is found using the G value calculated as in Eq. (5) [3]:

$$G = -2\ln\left(\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right) \tag{5}$$

The calculated G value shows a chi-square distribution with (p-1) degrees of freedom [3].

The data used in this study were generated by using a simulation program written in FORTRAN supported by Microsoft Power Station Developer Studio and IMSL Library. In this simulation study, one dependent and four continuous independent variables that were randomly drawn from the standardized normal distribution were taken into consideration in the constructed binary logistic regression model.

First, half of the random numbers in 5 variables were randomly taken from the first population and the other half was randomly drawn from the second population, having the standard normal distribution. Then, the random numbers in the dependent variable were encoded as 1 and 2 according to the population from which they were randomly taken.

In the calculation of test power, while the mean of the first population (μ_1) was not altered, the mean of the second population (μ_2) was increased by adding different amounts as standard deviation (δ). In this way, the differences (δ) between the population means from which the independent variables were taken against the codes of 1 and 2 in the dependent variable were 0.5, 1.0, 1.5 and 2.0 standard deviation (σ).

To calculate the type I error and test power, the simulation study was performed 100000 times and repeated for the sample sizes of 10, 20, 30 and 40. To inspect the effect of multicollinearity, which is defined as a high correlation among independent variables, the various degrees of multicollinearity, namely 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% were taken into consideration.

The type I error rate was calculated by dividing the number of simulation runs in which the control hypothesis that the constructed binary logistic regression model is not statistically significant was rejected by the total number of simulation runs. The type I error rate was determined as 5% at the beginning of the study.

Results

The type I error rates

First, the simulation was run 100000 times under different degrees of multicollinearity and sample sizes when the values of the dependent and independent variables were taken from the two populations having the same mean ($\mu_1=\mu_2$). Then, the fulfilled type I error rate was calculated for all the cases. The calculated rates of type I error for different sample sizes were plotted against the multicollinearity degrees (Figure 1).

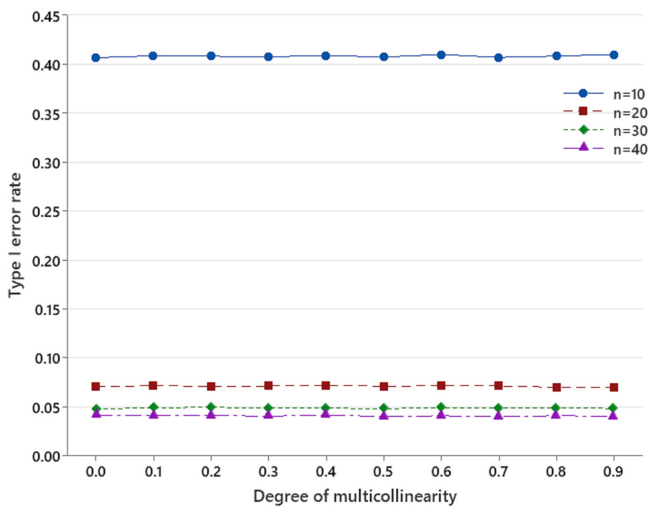

Figure 1. The type I error rates observed from 100000 simulation runs under different degrees of multicollinearity and sample sizes

As shown in Figure 1, increasing degrees of multicollinearity did not affect the type I error rate for all sample sizes. In other words, the type I error rate achieved when there was no multicollinearity was almost the same as that obtained under 90% of multicollinearity.

The attained type I error rate at the end of simulations runs was virtually very close to the predetermined type I error rate for the sample size of 30, which was not affected by changing degree of multicollinearity. There were considerable discrepancies between the attained and predetermined rates of type I error for the sample sizes of 20 and 40.

On the other hand, the small sample sizes in binary logistic regression analysis resulted in an abrupt increment in the type I error rate, as shown by the fact that the calculated rates for the sample size of 10 were unexpectedly higher than the predetermined type I error rate (Figure 1).

The power of binary logistic regression models

The powers of the binary logistic regression under all the above-mentioned cases were calculated for 0.5, 1.0, 1.5 and 2.0σ of the differences (δ) between the population means and were plotted against the degrees of multicollinearity in Figures 2, 3, 4 and 5, respectively.

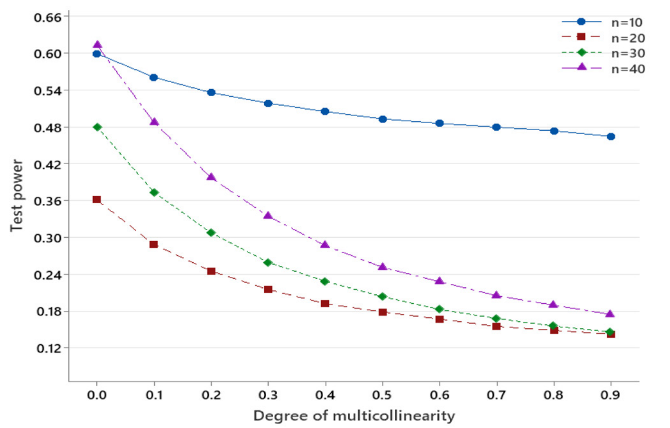

Figure 2. The power of the test observed from 100000 simulations runs under different degrees of multicollinearity and sample sizes when δ=0.5 (μ_1: μ_2=0.0:0.5)

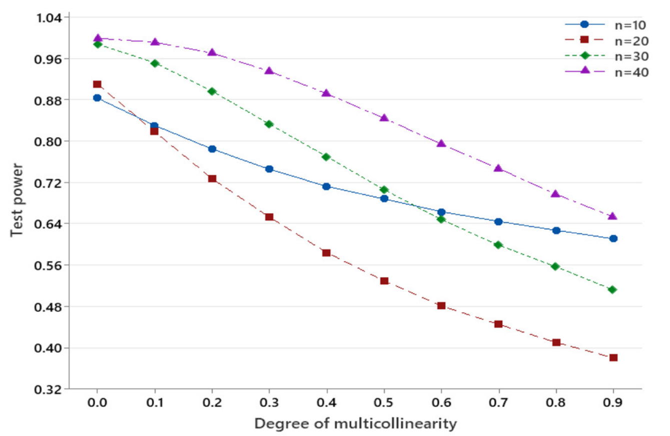

Figure 3. The power of the test observed from 100000 simulations runs under different degrees of multicollinearity and sample sizes when δ=1.0 (μ_1: μ_2=0.0:1.0)

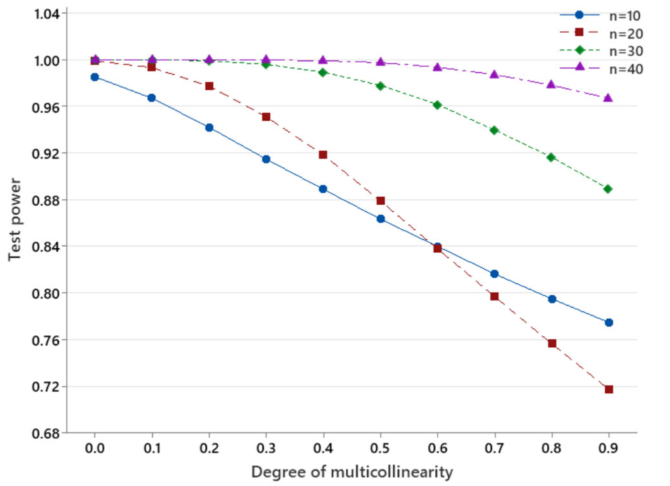

Figure 4. The power of the test observed from 100000 simulations runs under different degrees of multicollinearity and sample sizes when δ=1.5 (μ_1: μ_2=0.0:1.5)

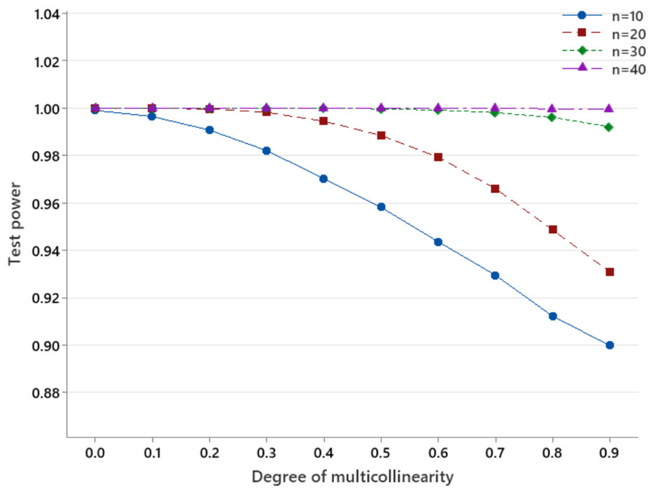

Figure 5. The power of the test observed from 100000 simulations runs under different degrees of multicollinearity and sample sizes when δ=2.0 (μ_1: μ_2=0.0:2.0)

As seen in Figure 2, the power of the binary logistic regression for δ=0.5 was least affected by the increasing degrees of multicollinearity when the sample size is 10. There was a 22.46% decline in the test power when the multicollinearity was raised from 0% to 90%, which cannot be accepted as reliable due the fact that the sample size of 10 is not sufficient for the binary logistic regression model including 4 independent variables, as stated in the literature [16,18]. However, as it can be seen from Figure 2, there was a sharp decline in the power of binary logistic regression analysis with increasing degree of multicollinearity for the other sample sizes. An abrupt decline, being 71.50%, in the test power was observed for the sample size of 40 when the multicollinearity was increased from 0% to 90%. The calculated test power verified that the power of binary logistic regression increased when the means of two populations sufficiently differentiated from each other (δ=1.0), as exhibited by the fact that all the calculated test powers were very close to 100% under all cases (Figure 3). On the other hand, the fact that there was a marked decline in test power with escalating multicollinearity for all the sample sizes clarified that the binary logistic regression was the most powerful if there was no strong multicollinearity among independent variables when there was a 1.0σ difference between population means (Figure 3). The most drastic decline in test power, which was 58.21%, was acquired when the sample size equals 20. The calculated power of the test under all the circumstances for δ= 1.5 and δ=2.0 were plotted against the degrees of multicollinearity in Figures 4 and 5, respectively. As seen in Figures 4 and 5, if the two populations were satisfactorily separated from each other, the binary logistic regression was an acceptably powerful test. It was, however, noticed that there was a slightly downward trend in test power with increasing multicollinearity. As can be noticed in Figures 4 and 5, the negative impact of the rising degrees of multicollinearity on the test power can be avoided if the number of observation is sufficiently large, for example, more than 40 for δ=1.5, and 30 or more than 30 for δ=2.0.

Discussion

The effects of multicollinearity on type I error rate and test power of binary logistic regression models were discussed in light of the relevant literature. No study in which both were investigated together has been found in the available literature.

In our study, the small sample sizes had a serious effect on the type I error rate in binary logistic regression models. The calculated rates from the sample size 10 were unexpectedly higher than the predefined type I error rate. As a matter of fact, the literature review found studies that confirmed this finding and emphasized the importance of sample size for logistic regression analysis [13,15-18]. They reported that the ratio of the number of observations in each variable to the number of independent variables should be equal to 10 or more than 10 to obtain the strongest maximum likelihood estimators from the multivariate models.

In a similar study, the type I error rates for a binary logistic regression model were examined at different levels of multicollinearity. The results of the study were found to be similar to our study. The study mentioned that multicollinearity may have a limited effect on the type I error rates of the binary logistic regression model but emphasised that these effects should not be ignored [19].

Our study found that the power of binary logistic regression models declined sharply with increasing levels of multicollinearity at all sample sizes. This was more pronounced for small sample sizes. The negative effect of multicollinearity in binary logistic regression models also increased as the differences in standard deviation units created between random variable populations to calculate power increased. The results of the study were also supported by the literature [16,18].

Conclusion

The following conclusions can be drawn depending on the

DOI: 10.5455/medscience.2023.08.146

findings of this study:

1. The type I error rate is not influenced by the multicollinearity as expected. However, it increased if the sample size is not sufficiently large. According to the results, the sample size should not be smaller than 30.
2. The binary logistic regression is not a powerful test if the populations from which the observations come are not adequately separated from each other.
3. The test powers calculated for varying degrees of multicollinearity and different sample sizes clarified that if the two populations were satisfactorily alienated from each other and if the number of observations was adequately large, multicollinearity was not a severe drawback in the binary logistic regression analysis.
4. In general, the number of observations to be used in binary logistic regression should not be less than 40 to avoid the negative impact of multicollinearity, if exists.

Based on the findings of our study, we think that multicollinearity should be considered and eliminated when necessary depending on the sample sizes in logistic regression analysis as well as in linear regression analysis.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
The study is a statistical theorem study and Ethics committee permission is not required. No research has been conducted on any living creature (human and animal).

This study was produced from the master's thesis conducted by the first author under the supervision of the second author at Ankara University. Part of it was presented as an oral presentation at the 13th International Symposium on Econometrics, Operations Research and Statistics held between 24-26 May 2012.

References

1. Kutner M, Nachtsheim C, Neter J. MP. Applied linear regression models-revised edition with student CD. 4th edition. NY: McGraw Hill/Irwin, 2004;701.
2. Draper NR, Smith H. Applied regression analysis. 3rd edition. New York, USA: John Wiley & Sons, 2014;736.
3. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd edition. New. York, NY: John Wiley & Sons, Inc; 2000;373.
4. Asar Y. Some new methods to solve multicollinearity in logistic regression. Communications in Statistics-Simulation and Computation. 2017;46:4,2576-86.
5. Berkson J. Application of the logistic function to bio-assay. Journal of the American Statistical Association. 1944;39:357-65.
6. Berkson J. A statistically precise and relatively simple method of estimating the bio-assay and quantal response, based on the logistic function. Journal of the American Statistical Association. 1953;48:565-99.
7. Berkson J. Maximum likelihood and minimum $\mathcal{X}^2$ estimates of the logistic function. Journal of the American Statistical Association. 1955;50:130-62.
8. Cornfield J. Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Federation Proceedings. 1962;21:58-61.
9. Cox DR, Snell EJ. The analysis of binary data. 2nd edition. Chapman and Hall, London, 1989;236.
10. Halperin M, Blackwelder WC, Verter JI. Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. J Chronic Dis. 1971;24:125-58.
11. Finney DJ. Probit analysis. Cambridge, Cambridge University Press. 1971;333.
12. Aguilera AM, Escabias M, Valderrama MJ. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics and Data Analysis. 2006;50:1905-24
13. Hosmer DW, Lemeshow S, Sturdivant R. Applied logistic regression, 3rd ed. New York: Wiley, 2013;528.
14. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics. 2010;13:253-67.
15. Senaviratna NAMR, Cooray TMJA. Multicollinearity in binary logistic regression model. Theory and Practice of Mathematics and Computer Science. 2021;8:11-9.
16. Schmitz PIM, Habbema JD, Hermans J. The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. Statist. Med. 1983;2:199-205.
17. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373-9.
18. Pituch KA, Stevens JP. Applied multivariate statistics for the social sciences: analysis with SAS and IBM'S SPSS. 6th edition. New York, NY: Routledge: Taylor & Francis Group, 2016;816.
19. Ghaner AW. Repercussions of multicollinearity in binary logistic regression. Master thesis. Shawnee State University, USA, 2022.