Summer 2022

Repercussions of Multicollinearity in Binary Logistic Regression

Aaron Wayne Ghaner

SHAWNEE STATE UNIVERSITY

Repercussions of Multicollinearity in Binary Logistic Regression

A Thesis

By: **Aaron Wayne Ghaner**

Department of Mathematical Sciences

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Mathematics

Date: 7/27/2022

Accepted by the Graduate Department

Graduate Director, Date

The thesis entitled '**Repercussions of Multicollinearity in Binary Logistic Regression**' presented by **Aaron Wayne Ghaner**, a candidate for the degree of **Master of Science in Mathematics,** has been approved and is worthy of acceptance.

7/27/2022	
DATE	GRADUATE DIRECTOR
28 JUL 2022	
DATE	STUDENT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The type I error rates for the binary logistic regression model were examined across varying levels of multicollinearity. Population data sets were created using the statistical software package R and then used to create data suitable for binary logistic regression models. The results showed the type I error rates did not differ across multicollinearity levels, the percentage of accurately classified cases were unaffected, and the variance inflation factors were affected by more than just correlation between the independent variables. These results show that multicollinearity may have limited effects on the type I error rates of the binary logistic regression model; however, these effects should not be ignored.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Douglas Darbro, for guiding me away from all of the rabbit holes I would be determined to go down. I would also like to thank Alyssa, my better half, for supporting me through my research.

1.0 CHAPTER 1: INTRODUCTION

Binary Logistic Regression (BLR) has been used heavily in many fields such as education, medicine, economics, and engineering, to name a few. Any situation where a researcher needs to predict or model a binary outcome, such as pass/fail, present/not present, or diseased/not diseased, they will most likely go to BLR. When mixed with multicollinearity (MCL), the BLR model suffers some of the same issues as any other regression technique, such as inflated errors for the estimated parameters of the model, which leads to unstable parameter estimates. The presence of MCL in regression models has been studied heavily over the years; however, there is a lack of knowledge in what happens to the type I error rates of a BLR model when MCL is present. This study seeks to provide insight on the topic.

1.1 Background of the Problem

Binary Logistic Regression (BLR) is often the go-to statistical model when the dependent variable is dichotomous. It has a lot more ease of use than discriminant analysis and other models for its less strict assumptions. These assumptions are: independent error terms, linearity in the logit, no influential outliers, and no multicollinearity (MCL) present in the regressor variables (Tabachnick & Fidell, Classification of Cases, 2013). The independent error terms assumption is the idea that the predictor variables should come from unrelated cases; otherwise, the type I error rates will inflate. The next assumption listed was linearity in the logit, which implies any continuous predictor variable needs to have a linear relationship with the logit-transformation of the dependent variable. Continuing with the subsequent assumption listed above, there need to be no influential outliers in the data; outliers can lead to the model having poor predictive capabilities. Finally, the last assumption is no MCL, which means the predictors cannot have a linear relationship; otherwise, the BLR model will suffer the same consequences as any other regression technique.

MCL has been an ever-foreboding presence in statistical models. There is a reason there are numerous studies that have looked into the consequences of MCL. That reason is its potentially detrimental effects on the regression model. Just as with other regression techniques, having MCL in a BLR model can increase the standard errors of the model coefficients. Furthermore, the actual values of the predictors can be relatively untouched, meaning the reliability of the estimated parameters become unstable (Midi & Rana, 2013; King, Binary Logistic Regression, 2008; Schisterman, Perkins, Mumford, Ahrens, & Mitchell, 2017). Hosmer et al. (2013, pp. 145-150) went further and showed that not only can the standard errors increase, but the parameter estimates can also be affected. These findings indicate that MCL does not always have the same effects in each case. One case could see large error terms for the coefficients, and another could see the same behavior with widely different parameter estimates from what previous research has shown. A more recent finding, by Lieberman and Morris (2014), conducted a study comparing different binary classification methods when MCL was involved. They found that MCL was a relatively benign issue when only considering the predictive accuracy of the BLR model.

Researchers have shown that MCL may also indicate separation issues in the model regressors in more recent studies. Namely, Zeng and Zeng (2019) showed that perfect MCL in BLR implies quasi-complete separation, meaning a single variable almost separates the dependent variable. This same study also showed that MCL could imply complete separation, meaning that MCL isn't necessarily the only issue for some cases. The relation between MCL and separation is not the focus of this study, but it is worth keeping in mind when dealing with MCL or separation since one can imply some form of the other.

To sum up, many studies have focused on finding MCL in the BLR model and alleviating it, such as Senaviratna and Cooray's (Multicollinearity in Binary Logistic Regression Model, 2021). Similar numbers of studies have looked into what happens to the standard errors and coefficients of the BLR model when MCL is present (Midi & Rana, 2013; Hosmer, Lemeshow, & Sturdivant, Numerical Problems, 2013). Overall, there do not seem to be many studies that examine what occurs with type I error rates when MCL is present in a BLR model. This study seeks to address this lack of knowledge.

1.2 Statement of the Problem

Any interference in the capability of a BLR model to produce valid estimated parameters can lead to estimates that do not align with previous research, which leads to parameters appearing to be significant when in fact, they are not, or vice-versa. MCL is something very real that can cause such interference and is always present in real-world data, meaning it is hard to get away from (unless the experiment is very controlled). This is why many studies have examined what having MCL in the predictor variables means for the statistical model of choice and how to work around it if it is even an option to do so. Suppose the researcher's goal is to determine valid estimated parameters. In that case, there is not much literature on what MCL does to a BLR model with respect to the type I error rates for those predictors. This study aims to investigate just that.

1.3 Purpose

This study will be exploratory in nature and will use randomly generated data to determine the impact of MCL on the type I error rates of Binary Logistic Regression (BLR). Generating data is a practice used frequently by studies of this nature (Hosmer, Lemeshow, & Sturdivant, Numerical Problems, 2013; Lieberman & Morris, 2014; Peduzzi, Concato, Holford, & Feinstein, 1996; Brunner & Austin, 2009). This ensures there will be no errors in data entries or missing data entries and also allows for the creation of a hypothetical population with the desired parameters; most importantly, it is easily repeatable. Since there are no distribution assumptions when dealing with BLR, aside from the errors being binomial, the independent variables can be generated using a uniform distribution and not affect the analysis or scope of this study.

The independent variables X_1 and X_2 will be created using a method in R called *runif* that generates uniform random variables using the given parameters of size, min, and max, where size will determine the number of variates to create, min is the lower value for the interval, and max is the upper value for the interval. The remaining independent variables, X_3, X_4, and X_5, will be created by adding a uniform variate to X_1. This will allow for the level of correlation between variables in each sample to be controlled. Then a linear combination for each populations independent variables will be created and be used with the logistic function to create probabilities. These probabilities will then be used to create the dependent variables used in each population.

The design of this study will involve creating four population data sets, which will be referred to as P_1 to P_4. The size of each population will be 25,000. Each population will have one dichotomous dependent variable and two independent variables. Where each population will differ is the amount of correlation between their independent variables as well as the associations between the independent and dependent variables. The correlations will be 0.0 for P_1, 0.4 for P_2, 0.99 for P_3, and 0.999 for P_4. There will be 1000 samples of 500 drawn from each population, and a BLR model will be created from each of these samples. For each model created from a sample, a type I error for the model overall will be defined by whether any of the confidence intervals for any of the estimated parameters did not contain the population parameter.

1.4 Significance of this Study

Previous studies have examined MCL and its effects on the estimated parameters and their standard errors in many settings, such as linear regression (Adeboye & Olatayo, 2014; Brunner & Austin, 2009), BLR (Gujarati, Porter, & Pal, The Nature of Multicollinearity, 2009; Hosmer, Lemeshow, & Sturdivant, Numerical Problems, 2013; King, Binary Logistic Regression, 2008; Midi & Rana, 2013). Other studies have examined the effects of MCL on type II error rates of statistical equation modeling (Grewal, Cote, & Baumgartner, 2004). However, there is a lack of studies that focus on MCL when it comes to type I error rates for a BLR model. The type I error rates ultimately have an effect on the inferences about a sample's population, so knowing how MCL could affect this is useful knowledge.

1.5 Primary Research Questions

How does multicollinearity affect the type I error rates of a Binary Logistic Regression model?

How will the percentage of accurately classified cases for the model be affected when multicollinearity is present?

Will the Variance Inflation Factors increase as the correlation between predictor variables increases?

1.6 Hypotheses

As MCL increases in the model, the type I error rates will be perturbed slightly, if not at all.

The percentage of accurately classified cases will remain relatively stable.

The VIFs will increase as the correlation between predictor variables increases

1.7 Research Design

This study is exploratory in nature, and its design is relatively straightforward. The chosen instrumentation was the statistical software package R. Only one built-in method, *runif,* was needed for the creation of the population data sets. The overall goal of this study was to see how type I error rates compare when differing levels of MCL are present in a BLR model, so it is natural to wonder how the type I error rates will be compared among the different populations. This will be tackled by looking at every sample from each population, which will produce estimated parameters. Confidence intervals will be made for each estimated parameter, which will then be compared to the population parameters to determine if MCL affected the type I error rates. Since the BLR models are being looked at overall, a model was said to have a type I error if any of the estimated parameters had a type I error.

1.8 Conceptual Framework

There are many regression techniques used in education research. One of the more widely used techniques is Binary Logistic Regression (BLR). It allows for the prediction of group classification and has more lenient assumptions than its counterpart

Discriminant Analysis. BLR is very comparable to Linear Regression; it is a transformation of the Linear Regression model after all; however, the dependent variables in each model differ significantly, and the assumptions are less stringent. Instead of a continuous outcome variable, BLR has a dichotomous outcome variable; that is, the outcome can take on two values, which are mapped to 0 or 1 (e.g., pass/fail, alive/dead, infected/not infected, etc.). Many consider the assumptions for BLR to be easy to work with since they do not determine what type of distribution the independent variables come from and do not require equal variances in the predictors, as is the case in linear regression. What the assumptions for BLR do state are: the error terms are independent (each case needs to be independent of the others), the continuous independent variables are linearly related to the logit, there is a lack of influential outliers, and there's an absence of multicollinearity (MCL) (Tabachnick & Fidell, Classification of Cases, 2013).

For BLR, there are a couple of ways to determine if the model is a good fit and/or accurate. A log-likelihood value is one such method when producing a model. This value is the sum of a model's individual log-likelihood values, which are found for each data point (Tabachnick & Fidell, Classification of Cases, 2013). This value is then compared to the log-likelihood value of a model containing no predictors and a χ^2 test can be used to determine if the model with added predictors is statistically significant compared to the intercept-only (or null) model. Another method used to assess a model's goodness-of-fit is the Hosmer-Lemeshow test, which is another statistical test and has the null hypothesis of "the model fits the outcomes effectively," which means a good model would not have statistical significance for this test. As for a model's accuracy, contingency tables are used to help with understanding a model's validity. A contingency table entails comparing the observed outcomes to the model's predicted outcomes (Peng, Lee, & Ingersoll, 2002). Values called specificity and sensitivity can also be calculated from such a table, where specificity is the percentage of correctly classified 0's as 0's and sensitivity is the percentage of correctly classified 1's as 1's (Tabachnick & Fidell, Classification of Cases, 2013). The type I error rates being considered for this study refer to the type I error rates for the estimated predictors. A type I error rate is defined as "the rate of rejecting a true null hypothesis". If the rate is higher than the standard 5%, then the parameter estimates for the model are considered to be unreliable.

MCL is a linear relationship between one or more regressors that can be explained by single or multiple regressors (Gujarati, Porter, & Pal, The Nature of Multicollinearity, 2009). MCL is almost guaranteed to occur in real-world data, especially if that real-world data was not collected in a controlled experiment, which is usually not the case when it comes to education research. This leads to quite the need for determining if MCL is present. Some of the most widely used methods to find MCL in data are; the correlation matrix, the Variance Inflation Factor (VIF) for each variable, and/or the condition numbers and variance-component proportions (Belsley, Kuh, & Welsch, Detecting and Assessing Collinearity, 1980). Multiple detection methods exist due to the nature of how MCL can arise. The correlation matrix can show little to no collinearity between variables even though they could be highly correlated (Midi & Rana, 2013). This is where the more robust VIF comes into play, where a VIF of greater or equal to 5 is usually considered to indicate high MCL. The variance-component proportions are another more robust method than the correlation matrix. This method consists of finding singular values, which are then used to find how much variance each variable contributes to another and itself. Besley et al. even state that the variance-component proportions are better than the VIF due to the ability to determine the effects of MCL when one variable is explained by more than just one other variable (Detecting and Assessing Collinearity, 1980). As for how MCL can affect the BLR model, Stoltzfus (Logistic Regression: A Brief Primer, 2011) notes the BLR model is contorted by MCL via inflation of the standard errors for the regression coefficients, which others have also found (Hosmer, Lemeshow, & Sturdivant, Numerical Problems, 2013; Midi & Rana, 2013). However, this is not the only effect MCL can have. Hosmer et al. (2013) have shown that MCL can also negatively impact the estimated coefficients themselves.

As stated previously, the goal of a BLR model is to predict group classification. The classification of a case is ultimately determined by what the estimated parameters for the model are, and to determine if these estimated parameters are valid, inference procedures need to be used. These inferences entail finding whether the estimated parameters are statistically significant and finding each estimated parameter's 95% confidence interval.

The type I error rates for the estimated predictors are determined by the choice of α used, where an α of 0.05 would mean a type I error is expected 5% of the time and is a typical value used. If MCL has an effect on the type I error rates for the predictors, then the estimated predictors should differ from the population parameters with a different percentage than the set α level, meaning the number of times the confidence intervals for the model’s estimated predictors will not contain the value of the population parameter will differ compared to when MCL is not present. This means the type I error rates of the estimated predictors can be determined by checking for the inclusion of the population parameters within the 95% confidence intervals for each estimated parameter.

1.9 Assumptions, Limitations, and Scope

The results of this study are not limited by the distributions used for the independent variables; however, they are limited due to the variables being randomly generated, which limits the results to some very specific use cases. This study also only focused on the “simpler” cases of MCL. As in, only two independent variables are linearly related to one another. Due to how the type I error for a binary logistic regression model was defined, the type I errors could only be compared amongst the populations instead of with a 5% proportion of type I errors. Time constraints limited analysis of the percentage of accurately classified cases and variance inflation factors to the distributions of each.

1.10 Definition of Terms

BLR – Binary Logistic Regression

CI – Confidence Interval

MCL – Multicollinearity

PAC – Percentage Accurately Classified

1.11 Summary

This chapter briefly covered the main topics found in this study, which showed that current research has not yet covered how MCL affects type I error rates in BLR. The stage was also set for the methods to be used in determining the results of this study. The background of the problems at hand, as well as precise statements for each problem being investigated within the scope of this paper were also presented.

2.0 CHAPTER 2: LITERATURE REVIEW

2.1 Background of Binary Logistic Regression (BLR)

BLR is a statistical model used to determine group classification. It has been used quite heavily over the years and has risen in popularity more than its predecessor discriminant analysis. The reason for this seems to be due to its less strict assumptions, ease of use when interpreting the model coefficients, and the fact many social sciences have dichotomous dependent variables.

The idea behind BLR is to model a discrete random variable that is dichotomous by using independent variables that are continuous, discrete, or a mix of these types of random variables (King, Binary Logistic Regression, 2008; Tabachnick & Fidell, Practical Issues, 2013). The method of BLR can be derived from the Simple Linear Regression (SLR) model by considering the effects of the dependent variable being dichotomous (Kutner, Nachtsheim, Neter, & Li, Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005). The SLR model can be seen below in equation 2.1 with dependent variable Y_i, independent variable X_i, parameters β_i, and the error terms ε_i.

$$(\mathbf{2.1})\ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Analyzing the SLR model above when the Y_i are dichotomous $(Y_i = 0{,}1)$ as described by Kutner et. al. (Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005) means the expected values for the Y_i , after noting $E(\varepsilon_i) = 0$, will be $E(Y_i) = \beta_0 + \beta_1 X_i$.

If the Y_i are Bernoulli random variables, then we know the probability of $Y_i = 1$ is π_i and the probability of $Y_i = 0$ is $1 - \pi_i$. This means the expected value of the Y_i will be the probability of $Y_i = 1$ when the independent variable is X_i. However, SLR assumes normal error terms, but this assumption is violated for a dichotomous Y_i since the error terms are either $\varepsilon_i = 1 - (\beta_0 + \beta_1 X_i)$ or $\varepsilon_i = -(\beta_0 + \beta_1 X_i)$, when $Y_i = 1$ and $Y_i = 0$, respectively (Kutner, Nachtsheim, Neter, & Li, Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005). Another issue is the error variance is not constant since the variance for any particular level of X_i will depend on X_i, which can be seen in equation 2.2 below (Kutner, Nachtsheim, Neter, & Li, Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005).

$$(\mathbf{2.2})\ \sigma(\varepsilon_i) = E(Y_i)[1 - E(Y_i)] = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$$

One of the last issues with a dichotomous response variable in SLR is the model's output not being bounded between 0 and 1 (Kutner, Nachtsheim, Neter, & Li, Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005). This is a necessity when the response variable is dichotomous since the response itself will be a probability. To fix these issues the BLR model was created.

The BLR model can be seen below in equation 2.3.

$$(\mathbf{2.3})\ P(Y_i = 1|X_i) = \frac{e^u}{1 + e^{\mathrm{u}}}$$

where the Y_i are independent Bernoulli random variables, and u is:

$$(\mathbf{2.4})\ \mathrm{u} = \beta_0 + \sum_{j=1}^{n} \beta_j X_j$$

which is just the Linear Regression model. The predictors for the BLR model in equation 2.3 are found using a method called Maximum Likelihood (ML). This method seeks to find estimates for the parameters of the model by maximizing the probability of finding the data that was used; in other words, it maximizes the loglikelihood function for the data.

2.2 Assumptions of Binary Logistic Regression

The assumptions that need to be met for BLR to be most effective are; a linear relationship between the logit and any continuous predictors, independence of error terms, no influential outliers, and no multicollinearity (MCL) (Tabachnick & Fidell, Logistic regression, 2013). These were mentioned in the previous chapter, but are now discussed in more detail here.

The first assumption stems from the transformation used in creating the BLR model and can be tested using the Box-Tidwell test, which is a method of including interaction terms between each predictor and its natural logarithm in the BLR model. If any of these terms have statistical significance, then the assumption is violated and a transformation of them may alleviate the issue (Tabachnick & Fidell, Logistic regression, 2013). Other techniques to test for linearity in the logit are discussed by Hosmer et al. (Methods to Examine the Scale of a Continuous Covariate in the Logit, 2013), which are smoothed scatter plots, design variables, fractional polynomials, and spline functions.

Having independence of error terms means each case is unrelated to the others. In the event this assumption is violated the type I errors can become inflated due to overdispersion, which is a phenomenon where the cell frequencies vary far more than expected (Tabachnick & Fidell, Logistic regression, 2013). Ways to accommodate violating this last assumption are to use a categorical dependent variable within multilevel modeling so the interdependencies would be considered as part of the model or to perform transformations on the standard error terms (Tabachnick & Fidell, Logistic regression, 2013).

As for influential outliers, these can lead to poor model fit and can be checked by examining the residuals. One method of examining residuals is to standardize them, any residuals that can be considered as “large” should be removed from the model to produce a better fitting one (Tabachnick & Fidell, Logistic regression, 2013). Other methods noted by Kutner et al. (Logistic Regression, Poisson Regression, and Generalized Linear Models, 2005) use deviance statistics, Pearson χ^2 statistics, or Cook’s distances. The deviance and Pearson statistics are used to create deltas to determine how much each case influences the model fit, while Cook’s distance is a measure of the fitted values when the i^{th} case is deleted from the model. Currently, it is not always feasible to determine when one or more of these statistics presents an influential observation, instead plots are used to provide a better understanding of the situation.

The last assumption of no MCL in the predictor variables means no high amounts of MCL should be present between the predictor variables and is generally considered to be when correlations between variables are 0.8 or higher. The violation of no MCL can lead to large standard errors, inconsistencies in the estimated predictors, and possibly cause the solution for the model to not converge (Tabachnick & Fidell, Logistic regression, 2013; Hill & Adkins, 2001; King, Collinearity, 2008).

Another criterion for this type of model deals with the ratio of cases to variables. Having too little a ratio can produce similar effects to that of having MCL. This is why the recommended ratio is to have at least 10 cases per variable (Peduzzi, Concato, Holford, & Feinstein, 1996).

2.3 Multicollinearity (MCL) and its Effects on Regression Techniques

Data collected outside of a well-controlled experiment will almost always have MCL. The phenomenon of MCL occurs when two or more predictor variables are linearly dependent. This dependency can be shared between multiple variables or only between two of them. MCL is so prevalent in statistical analysis alone that many methods have been developed to work around it. Some of these methods for BLR are to increase the sample size, remove any redundant variables, perform a transformation on the variables, or do nothing and report the MCL issue (Gujarati, Porter, & Pal, The Nature of Multicollinearity, 2009; King, Collinearity, 2008; Belsley, Kuh, & Welsch, Detecting and Assessing Collinearity, 1980; Kleinbaum, Kupper, Nizam, & Rosenberg, 2014). Each of these methods have its limitations, though. Retrieving more data is not always viable or feasible, which leads into the other remedies. Removing redundant variables can lead to specification error, which can have worse consequences than MCL since it can lead to biased estimates. As for data transformations, depending on the type of transformation, the assumptions of the model can be violated (Gujarati, Porter, & Pal, The Nature of Multicollinearity, 2009).

The actual issue of MCL in the BLR model arises in how the parameters for the model are created. The maximum likelihood (ML) method is not immune to MCL, rather it is just as susceptible as Ordinary Least Squares is for Simple Linear Regression. The main problem comes from the Jacobian matrix used in the calculations for the estimated parameters (Hill & Adkins, 2001). This matrix needs to be inverted, and when a matrix is ill-conditioned, as described by Belsley et al. (Detecting and Assessing Collinearity, 1980), this leads to massive values when dividing by a number close to zero, which propagates to the parameter estimates and their errors, or even leads to the ML method not being able to converge (Hill & Adkins, 2001; Tabachnick & Fidell, Multicollinearity and Singularity, 2013).

So far, this study has mainly labeled MCL as an issue with the predictors themselves; however, this is not always the case. When a population is sampled, there is sample variation. This variation can mean one sample has no MCL, but another does. This is why one of the suggested remedies of fixing the issue of MCL is to get more sample data. For most researchers, this option is typically not feasible, so they need to work with what they have, or not do anything about it while at least reporting that MCL was present (Gujarati, Porter, & Pal, The Nature of Multicollinearity, 2009). Other methods of removing MCL from a model entail removal of the offending variable(s), combining them, or using a technique like principal component analysis (PCA) to create new predictors from the created components (Tabachnick & Fidell, Multicollinearity and Singularity, 2013). MCL creates redundancy, which makes it difficult to determine what variable describes what variation in the response variable, so the removal of the redundant predictor is one of the more straightforward approaches to correcting the issue. Combining the redundant variables is an option to keep each initial predictor in the model but still eliminates the MCL issue. The method of using PCA to determine the underlying structure of the variables allows for the initial data to be used while creating new predictors from the components produced to use in the model. However, if the researcher's goal is purely prediction, then some have suggested that MCL is not of any concern (Tabachnick & Fidell, Multicollinearity and Singularity, 2013; Lieberman & Morris, 2014).

2.4 Detection Methods for Multicollinearity and Ways to Work with/around it

Needless to say, many tools have been created to detect MCL. One of the most basic is the correlation matrix for a data set, which provides the coefficient of simple correlation for each pair of variables (Tabachnick & Fidell, Correlation, 2013; Kutner, Nachtsheim, Neter, & Li, Multicollinearity and Its Effects, 2005). However, MCL is not limited to two variables; meaning the coefficients of simple correlation within the correlation matrix can be misleading when multiple variables are at play (Kutner, Nachtsheim, Neter, & Li, Multicollinearity and Its Effects, 2005). This led to the development of other methods to determine the extent of MCL. One such method is the Variance Inflation Factor (VIF). This allows one to see how much the variance is inflated compared to when the predictor variables are not correlated at all. A VIF of 1.0 means there is no MCL and a VIF of 5 or even 10 or more is said to indicate the presence of MCL (Kleinbaum, Kupper, Nizam, & Rosenberg, 2014; Kutner, Nachtsheim, Neter, & Li, Multicollinearity Diagnostics - Variance Inflation Factor, 2005). While VIFs are useful, they are not infallible, since numerous accounts of MCL cannot be revealed by VIFs alone (Kutner, Nachtsheim, Neter, & Li, Multicollinearity Diagnostics - Variance Inflation Factor, 2005). Another method of determining the presence of MCL involves finding the eigenvalues (sometime referred to as principal components) of the correlation matrix for the predictor variables. A larger eigenvalue indicates a larger contribution to the variation in the predictors, on the other hand, a small (zero or nearly zero) eigenvalue indicates MCL is at play within the regressors (Kleinbaum, Kupper, Nizam, & Rosenberg, 2014). From these eigenvalues come condition indexes, the condition number (CN), and variance proportions. Each eigenvalue will have a condition index associated with it, which is found by taking the square root of the ratio between the largest eigenvalue, λ_{max}, and the j^{th} eigenvalue, λ_j. Equation 2.5 gives the mathematical representation for a condition index, CI_j.

$$\mathbf{(2.5)}\ CI_j = \sqrt{\lambda_{max}/\lambda_j}$$

The CN is the largest condition index obtained, meaning it is the ratio between λ_{max} and the smallest eigenvalue. The larger the condition index, the more of a problem MCL is for the data, and when the CN is 30 or more, further investigation of the data set may be needed (Belsley, Kuh, & Welsch, Detecting and Assessing Collinearity, 1980). Further investigation entails looking at the variance proportions. These proportions reveal the amount of estimated variance each regression coefficient has among the eigenvalues (Belsley, Kuh, & Welsch, Detecting and Assessing Collinearity, 1980; Kleinbaum, Kupper, Nizam, & Rosenberg, 2014). These variance proportions allow for analysis of the structure MCL has in the data set. Any variance components that contribute high proportions of variance to an eigenvalue with a high condition index indicates MCL. An example of such a case would be when a principal component has two or more proportions of 0.5 (Kleinbaum, Kupper, Nizam, & Rosenberg, 2014).

2.5 Type I Error Rates for the Estimated Predictors in Binary Logistic Regression

The type I errors (false positives), and by association type II errors (false negatives), that can occur when testing the validity of the estimated predictors in BLR (or any other model) are rooted in hypothesis testing. This is a practice that has come from the need to make an informed decision when dealing with statistical estimates. The general idea behind which is to assume two states for the statistic in question. The first is the state of the null hypothesis and the second is that of the alternate hypothesis. Each of these states are associated with a probability distribution for the test statistic of interest. For the BLR model these test statistics are derived from the model's estimated parameters.

To formulate a decision about which state is acceptable based on the statistical evidence an α level must be set before investigating the hypothesis. This α level is the probability of committing a type I error (also known as the level of significance) that is considered acceptable. A typical α level is .05; however, sometimes other levels, such as .01 or .10, are used (Tabachnick & Fidell, Review of Univariate and Bivariate Statistics, 2013). Type I errors arise when the null hypothesis is true but is ultimately rejected due to the statistical evidence pointing to the alternate hypothesis as the more likely state. As for type II errors, these occur when the alternate hypothesis is true but the null hypothesis is accepted due to the statistical evidence being in its favor.

2.6 Random Number Generators, Random Sampling, and R

There are many fields that have use for Random Number Generators (RNGs). Especially statistics, such as studies using Monte Carlo methods. Their use stems from the need to create random samples of data, or at least to create observations (the output of the RNGs) that appear to be independent and identically distributed (i.i.d) uniform random variables. These values created by the RNG are usually treated as probabilities, which are then transformed into some arbitrary random variable (i.e., binomial, normal, exponential, etc.) by using the associated distribution function for this arbitrary variable (L'Ecuyer, 2012). One of the most widely used RNGs currently is the Meresenne Twister (MT) algorithm developed by Makoto Matsumoto and Takuji Nishimura (1998) and was revised in 2008 to correct some of the flaws found in the initial implementation for the algorithm (Saito & Matsumoto). This particular algorithm is used as the default RNG by many software packages, including the statistical software package R.

Despite their wide use, RNGs are not without their limitations. Some limitations of RNGs lie within the domain of computers, which use binary floating point as a way to represent numbers and only have finite amounts of resources (Knuth, 1998). This translates to computers not being able to accurately represent all real numbers and that they have the inability to produce infinite sequences. The workaround, for RNGs at least, is to produce integers between 0 and some value, say m, then dividing the integers by m to produce values between 0 and 1. The capacity for the value of m is known as the period of the RNG and depends on the algorithm. For the MT algorithm, its period is 2^{19937}-1. Meaning it can generate 2^{19937}-1 unique values before repeating one. The benefits of the MT algorithm, aside from its computational speed and performance, is its massive period and it being 623-equidistributed (Matsumoto & Nishimura, 1998). This last property translates to having equal amounts of values within each hypercube of the state space. On the other hand, every RNG has the flaw of being periodic and predictable, they are algorithms after all (L'Ecuyer, 2012). Although the limitations of RNGs are not always generalizable, each family has their own. One of the limitations of the F2 family, which the MT algorithm falls into, is they all fail the matrix rank test and the linear complexity test (L’Ecuyer & Panneton, 2009). Both of these tests are used to determine if a linear relationship exists between the observations produced (L’Ecuyer & Simard, 2007). This indicates there is bias in the observations created by the MT algorithm; however, this is a limitation in the MT algorithm’s inability to create observations that mimic a sequence that is genuinely random, and does not mean its output is invalid (L'Ecuyer & Panneton, 2009).

Random sampling is not far off of the concept of RNGs. Typically, the data that the sample is pulled from will be enumerated and then picked at random by way of RNGs. Instead of producing numbers within the set [0, 1], the RNG will produce integers from a given set. As an example, generating a random sample of three from the set of integers 1 to 10 would mean the RNG would be limited to using values between 1 and 10, then randomly picking three of them to produce the sample.

Despite the limitations of the MT algorithm, its issues will not be apparent in this study. The limitations arise when the number of observations created gets closer to the size of the RNGs period. This study will require a very small fraction of observations in comparison to the period of the MT algorithm.

3.0 CHAPTER 3: METHODOLOGY

This study aims to investigate the effect of multicollinearity (MCL) on the type I error rates of binary logistic regression (BLR). What follows is the methodology on what the structure of data was, how the data was generated, and how each hypothesis iterated in the first chapter was tested.

Four population data sets were created for this study; P_1, P_2, P_3, and P_4. Each population had one dichotomous response variable and two continuous predictors, where the population parameters were $\beta_0=0$ and $\beta_1=\beta_2=1$ for each population. The populations were created in a similar manner as described in the work of Hosmer et al. (1997) as well as Hosmer and Hjort's (2002).

The size of each simulated population was 25,000. Each population had two independent variables and a binary outcome variable. The independent variables were created using uniform variates. The X_1 variable was a uniformly distributed random variable on the interval [-6, 6] and was used in each population. This interval was chosen since Hosmer et. al. noted it to be reliable for getting mainly small and large probabilities (1997). The remaining independent variables for each population were X_2 for P_1, X_3 for P_2, X_4 for P_3, and X_5 for P_4. The X_2 variable shared the same creation method as X_1 and had no notable correlation with X_1. The X_3 variable was created by adding a uniform variate on the interval [-12.5, 12.5] to X_1, which made X_3 correlate with X_1 at about the 0.4 level. The X_4 variable was created the same way as X_3; however, the uniform variate was on the interval [-0.75, 0.75] to create a correlation of around 0.99 between X_1 and X_4. Finally, the creation of X_5 also mimicked that of X_3 but the added uniform variate was on the interval [-0.25, 0.25], which created a correlation with X_1 of around 0.999.

Linear relationships for the independent variables in each population were then created. These relationships were $Z_1=X_1 + X_2$ for P_1, $Z_2=X_1 + X_3$ for P_2, $Z_3=X_1 + X_4$ for P_3, and $Z_4=X_1 + X_5$ for P_4. To create the dependent variable for each population an independent uniform variate, u, on the interval [0,1] was compared to the true logistic probability, $\pi(Z_i)$ for i=1, 2, 3, 4. If $u \leq \pi(Z_i)$, then Y = 1; otherwise, Y = 0. This independent comparison of u and $\pi(Z_i)$ prevented separation issues within the data.

A power analysis was performed to determine the necessary sample size to achieve a power of 0.95. G*Power was used to determine the needed sample size. The test family was set to Chi-square tests, then the statistical test chosen was the Goodness-of-fit test for contingency tables, and the power analysis type was set to calculate the required sample size. The parameters for this type of power analysis were set as follows: the effect size was set to 0.1 in case the effects of MCL were low, α was set to .05, power was set to 0.95, and the degrees of freedom was set to 3 since there are four levels of MCL being compared. These values in G*Power produced a required sample size of 1,717. This indicates that the total sample size of 4,000 (1,000 samples per MCL level) for this study was sufficient to achieve a power of 0.95.

A total of 1,000 samples of 500 were drawn from each population. Before creating a BLR model for each sample, the assumptions of BLR were tested. The assumptions tested were: linearity in the logit, independence of error terms, and no influential outliers (Tabachnick & Fidell, Logistic regression, 2013). The linearity in the logit assumption was tested using the Box-Tidwell test. The error terms were assumed to be independent given the nature of how the data was created. The lack of influential outliers was tested by examining the standardized residuals for each model. If any residuals were found to be three standard deviations higher than the mean, then that case was marked as an outlier. The level of MCL for each population was verified by finding the VIF values of the variables and condition numbers. A VIF of 5 or more was considered to be an influential level of MCL and a condition number of 30 or more was also considered to mean high MCL levels.

After the model assumptions were verified a BLR model was created for each sample. Then 95% confidence intervals were made for all parameters within each model. If the confidence interval contained the population parameter, then that case was considered accurate. On the other hand, if the confidence interval for any estimated parameter did not include the population parameter, then that particular case was considered to be a type I error. The type I error rates for each overall model were determined by whether any of the estimated predictors had a type I error. So, if a sample from a population created a model where b_1 or b_2 did not fall within their respective confidence intervals, then a type I error occurred in the model.

If MCL did not have any effect on the type I error rates, then the type I error rate for each population will be the same; otherwise, at least one of the type I error rates differed. To determine this, a Chi-square goodness-of-fit test was used. The null hypothesis was set to “$er1 = er2 = er3 = er4$”, where *er* was the type I “error rate”, and the alternate hypothesis was “at least one type I error rates differed from the others”.

Due to time limitations, the Percentage of Accurately Classified cases (PAC) were only looked at from a distribution standpoint. The model PACs were calculated and from these, the descriptive statistics for their distributions were found and the PACs for the population model (the BLR model for the entire population data set) were compared to these distributions.

For similar reasons, the VIFs for each population were only examined from a distributional point of view as well. The VIFs for each model were created and the descriptive statistics for each model VIFs were found.

In summary, the design of this study focused on creating population data to then sample from. Each sample, if it passed the assumptions for BLR, was then used to create a BLR model. Inferences were then performed on each model's parameters. If the confidence interval for an estimated parameter did not contain the value of the population parameter, then it was marked as a type I error and the model was said to have a type I error overall. The type I error rates were then determined by adding up the errors across estimated parameters and dividing by the total number of models for that specific population. What follows are the results of this study.

4.0 CHAPTER 4: RESULTS

This chapter details all findings. More precisely, this chapter analyzes the results of each hypothesis listed in the first chapter of this study, which are listed below:

- As MCL increases in the model, the type I error rates will be perturbed slightly, if not at all.
- The percentage of accurately classified cases will remain relatively stable.
- The VIFs will increase as the correlation between predictor variables increases

A χ^2 goodness-of-fit test showed the type I error rates for binary logistic regression models did not change across varying multicollinearity (MCL) levels, $\chi^2(3) = 3.132$, p=0.372. One reason for this result can be seen in Table 1, specifically population P_4, where the standard errors were so large that the true parameter estimates were within the confidence intervals for the respective estimated parameters.

Table 1 Representative Samples Models from each Population

Sample	Parameter	Parameter Estimate	p	95% CI	R^2_{HL}
		P_1			
1	β_0 (S.E.)	0.113 (0.166)	.498	(-0.213, 0.442)	0.682
	β_1 (S.E.)	0.954 (0.097)	<.001	(0.778, 1.161)	
	β_2 (S.E.)	0.965 (0.099)	<.001	(0.786, 1.175)	
2	β_0 (S.E.)	0.247 (0.164)	.133	(-0.073, 0.573)	0.659
	β_1 (S.E.)	1.00 (0.102)	<.001	(0.819, 1.221)	
	β_2 (S.E.)	0.973 (0.102)	<.001	(0.789, 1.188)	

3	**β_0 (S.E.)**	-0.139 (0.169)	0.411	(-0.474, 0.191)	0.589
	β_1 (S.E.)	0.918 (0.093)	<.001	(0.751, 1.115)	
	β_2 (S.E.)	0.921 (0.092)	<.001	(0.755, 1.116)	
P_2					
1	**β_0 (S.E.)**	-0.017 (0.163)	0.918	(-0.337, 0.305)	0.750
	β_1 (S.E.)	0.845 (0.084)	<.001	(0.693, 1.022)	
	β_2 (S.E.)	0.703 (0.093)	<.001	(0.531, 0.897)	
2	**β_0 (S.E.)**	-0.125 (0.175)	0.475	(-0.473, 0.217)	0.711
	β_1 (S.E.)	1.030 (0.107)	<.001	(0.837, 1.260)	
	β_2 (S.E.)	0.946 (0.116)	<.001	(0.734, 1.192)	
3	**β_0 (S.E.)**	0.218 (0.183)	0.234	(-0.140, 0.580)	0.671
	β_1 (S.E.)	0.971 (0.099)	<.001	(0.792, 1.181)	
	β_2 (S.E.)	0.944 (0.119)	<.001	(0.727, 1.195)	
P_3					
1	**β_0 (S.E.)**	-0.045 (0.211)	0.832	(-0.462, 0.369)	0.734
	β_1 (S.E.)	0.814 (0.493)	<.10	(-0.139, 1.807)	
	β_2 (S.E.)	1.190 (0.570)	<.05	(0.098, 2.346)	
2	**β_0 (S.E.)**	-0.084 (0.214)	0.693	(-0.511, 0.333)	0.795
	β_1 (S.E.)	0.776 (0.502)	0.122	(-0.193, 1.788)	
	β_2 (S.E.)	1.304 (0.558)	<.05	(0.236, 2.437)	
3	**β_0 (S.E.)**	0.639 (0.219)	<.01	(0.224, 1.085)	0.818
	β_1 (S.E.)	1.107 (0.497)	<.05	(0.155, 2.114)	

	β_2 (S.E.)	0.962 (0.529)	<.10	(-0.57, 2.029)	
P_4					
1	**β_0 (S.E.)**	-0.024 (0.215)	0.912	(-0.448, 0.400)	0.800
	β_1 (S.E.)	2.011 (1.641)	0.220	(-1.148, 5.330)	
	β_2 (S.E.)	-0.123 (1.651)	0.940	(-3.407, 3.111)	
2	**β_0 (S.E.)**	0.032 (0.235)	0.891	(-0.432, 0.497)	0.796
	β_1 (S.E.)	0.748 (1.562)	0.632	(-2.316, 3.849)	
	β_2 (S.E.)	1.417 (1.639)	0.387	(-1.778, 4.693)	
3	**β_0 (S.E.)**	-0.530 (0.303)	<.10	(-1.164, 0.035)	0.784
	β_1 (S.E.)	9.619 (2.695)	<.001	(4.817, 15.488)	
	β_2 (S.E.)	-6.890 (2.450)	<.01	(-12.124, -2.418)	

One oddity of the data was within population P1, which was designed to have no correlation between the independent variables. However, the average VIF as 4.269 (0.852), which was higher than the average VIF of 1.767 (0.287) for population P2, which was designed to have a small amount of correlation between the independent variables (Table 2). This shows that correlation between two variables may not be all that is needed for MCL to be present, or that VIF values are not as reliable as other measures, such as condition numbers, which did not have this phenomenon occur (Table 3).

Table 2 Average VIFs by Population

	X_1	**X_2**	**X_3**	**X_4**	**X_5**
P_1	4.269 (0.852)	4.269 (0.852)	--	--	--
P_2	1.767 (0.287)	--	1.767 (0.287)	--	--

P_3	5.185 (1.267)		--	5.185 (1.267)	--
P_4	40.675 (10.897)	--	--	--	40.675 (10.897)

Table 3 Average Condition Numbers by Population

	P_1	**P_2**	**P_3**	**P_4**
CN	1.073 (0.032)	1.594 (0.071)	16.045 (0.448)	48.127 (1.392)

The samples from P1 had parameter estimates that aligned with the true population parameters, as did the samples from P2 (Table 1). However, the third and fourth populations had increased standard errors and slightly to massively erratic parameter estimates (Table 1).

The percent of accurately classified cases (PAC) for each sample model did not vary from the population model PAC (Table 4). The parameters for each population were set to $\beta_0=0$ and $\beta_1=\beta_2=1$.

Table 4 PACs for Population Models and Sampled Models

	P_1	**P_2**	**P_3**	**P_4**
Population PAC	0.896	0.913	0.938	0.941
Sample PAC	0.896 (0.013)	0.915 (0.012)	0.940 (0.011)	0.942 (0.010)

The distributions for the PACs were symmetrical about their means and the means appeared to have increased as the MCL levels increased (Figure 1).

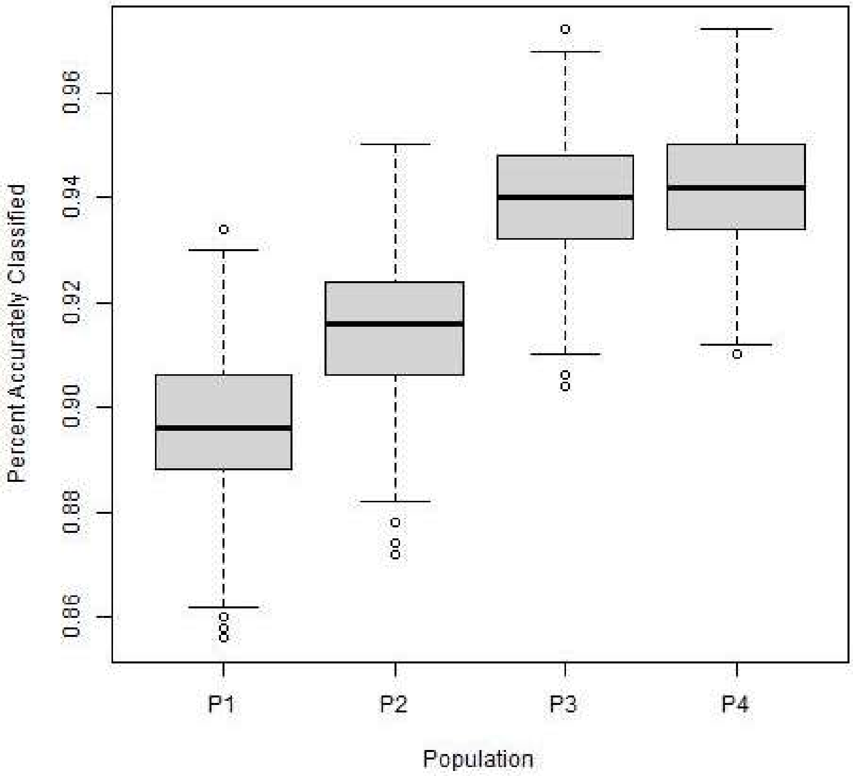

Figure 1 Percent Accurately Classified cases across Populations

Shapiro-Wilk’s tests showed each PAC population did not follow a normal distribution; P_1, W=0.995, p<.01; P_2, W=0.996, p<.05; P_3, W=0.995, p<.01; P_4, W=0.994, p<.001.

4.1 Population Data

After the BLR assumptions were verified the number of valid samples in each population were comparable; population P_1 had 983 valid samples, while P_2 had 986, P_3 had 948, and P_4 had 938.

The means and standard deviations for each predictor variable are shown in Table 5. Each independent variable was a uniform variate on [-6, 6], or at least created from one.

Table 5 Descriptive statistics for independent variables by population (* X_1 was used in each population)

	P_1	P_2	P_3	P_4
	mean (sd)	mean (sd)	mean (sd)	mean (sd)
X_1*	0.007 (3.459)			
X_2	0.007 (3.466)	--	--	--
X_3	--	0.003 (2.620)	--	--
X_4	--	--	0.006 (3.143)	--
X_5	--	--	--	0.007 (3.328)

The correlations for each X_i ranged from -0.009 to 0.999 (Table 6). It is worth noting that the purposefully created correlations are in the first row. That is, X_2 was designed to have little to no correlation with X_1, X_2 was designed to have a correlation with X_1 of around 0.4, and so on. The X_1 column and X_5 row were omitted due to them displaying redundant or non-meaningful information (Table 6).

Table 6 Correlations for the Independent Variables

	X_2	X_3	X_4	X_5
X_1	-0.009	0.430	0.992	0.999
X_2	1.00	-0.004	-0.010	-0.009
X_3	--	1.00	0.426	0.429
X_4	--	--	1.00	0.991

The proportion of 1's to 0's was consistent across populations, where each population size was 25,000 (Table 7).

Table 7 Proportions of 1's for the Dependent Variables

	P_1	P_2	P_3	P_4
Count of 1’s	12518	12545	12537	12504
% of 1’s	0.501	0.502	0.501	0.500

Significant positive correlations existed between the independent and dependent variables for each population and ranged from 0.542 to 0.846 (Table 8). These correlations increased as the correlation between the population variables increased and, by association, as the MCL levels increased (Table 8).

Table 8 Point-Biserial Correlations

	Y_i	cor	df	t	CI	p
X_1	Y_1	0.542	24998	101.94	(0.533, 0.551)	<.001
	Y_2	0.744	24998	176.26	(0.739, 0.750)	<.001
	Y_3	0.844	24998	248.39	(0.840, 0.847)	<.001
	Y_4	0.846	24998	250.59	(0.842, 0.849)	<.001
X_2	Y_1	0.542	24998	102.06	(0.534, 0.551)	<.001
X_3	Y_2	0.590	24998	115.46	(0.582, 0.598)	<.001
X_4	Y_3	0.841	24998	245.99	(0.838, 0.845)	<.001
X_5	Y_4	0.845	24998	250.23	(0.842, 0.848)	<.001

Overall, this chapter focused on the results found for this study, specifically the results associated with each hypothesis outlined in the first chapter. To summarize each result; MCL did not affect the type I error rates in BLR models, PACs did not vary from the overall population model PACs and had symmetric distributions that were not normally distributed, and the VIFs for each sample model did not increase as the correlation between the predictor variables increased. This first result follows findings by

5.0 CHAPTER 5: SUMMARY

This final chapter summarizes the results of the previous chapter, answers the hypotheses stated in the first chapter, and also includes recommendations for further research related to this study.

The focal point of this study was to determine the effects of multicollinearity (MCL) on type I error rates in the binary logistic regression (BLR) model. The results showed the varying levels of MCL had no effect on type I error rates and follow the trend outlined in the research by Grewal et al (Grewal, Cote, & Baumgartner, 2004), where the type II errors of structural equation models, which use regression techniques, were found to increase dramatically, meaning the type I errors would have decreased in turn.

The observed increased standard errors of the models when MCL was present agree with previous studies that focused on the effects of MCL on the parameter estimates and standard errors of the BLR model (Hosmer, Lemeshow, & Sturdivant, 2013; Midi & Rana, 2013; Stoltzfus, 2011; Adeboye & Olatayo, 2014). The percent accurately classified (PAC) for each model appeared to be unaffected by MCL, which is in line with the findings of Lieberman and Morris (The Precise Effect of Multicollinearity on Classification Prediction, 2014; Midi & Rana, 2013).

These results do not indicate the effects of MCL can be ignored. Instead, they point more towards how MCL is a very ingrained and intricate problem faced in regression techniques as a whole. However, MCL may not be as formidable a problem when researchers do not want to increase the type I errors and are only interested in classification prediction.

Possible expansions on this study could examine the effect of more problematic MCL, more problematic meaning multiple variables correlate with one another, or one could examine the type I errors of each estimated predictor within the models to have a finer detailed data set to determine if the type I errors were affected on a smaller scale or remained unchanged. Lastly, one could examine the effects of MCL when the population parameters differ across populations as well.

6.0 REFERENCES

Adeboye, O., & Olatayo, T. (2014). *Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients*. Retrieved October 10, 2021, from Research Gate: https://www.researchgate.net/profile/Olawale-

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Detecting and Assessing Collinearity. In D. A. Belsley, E. Kuh, & R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (pp. 85-169). Hoboken: John Wiley & Sons, Inc.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Detecting and Assessing Collinearity. In D. A. Belsley, E. Kuh, & R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (pp. 85-169). Hoboken: John Wiley & Sons, Inc.

Brunner, J., & Austin, P. (2009, March 30). *Inflation of Type I error rate in multiple regression when independent variables are measured with error.* Retrieved from Wiley Online Library: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.10004

David W. Hosmer, J., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression.* John Wiley & Sons, Inc.

Grewal, R., Cote, J. A., & Baumgartner, H. (2004, November 1). Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing. *Marketing Science*, 519-529. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.525.1921&rep=rep1&type=pdf

Gujarati, D. N., Porter, D. C., & Pal, M. (2009). The Nature of Multicollinearity. In D. N. Gujarati, D. C. Porter, & M. Pal, *Basic Econometrics* (pp. 315-345). McGraw Hill.

Gujarati, D. N., Porter, D. C., & Pal, M. (2009). Two-Variable regression: Interval Estimation and Hypothesis Testing. In *Basic Econometric* (pp. 111-122). McGraw Hill.

Hill, R. C., & Adkins, L. C. (2001). Collinearity. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics* (pp. 254-277). Blackwell Publishing.

Hosmer, D. W., & Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 2723-2738.

Hosmer, D. W., Hosmer, T., Cessie, S. L., & Lemeshow, S. (1997). A COMPARISON OF GOODNESS-OF-FIT TESTS. *Statistics in Medicine*, 965-980.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Numerical Problems. In D. W. Hosmer, S. Lemeshow, & R. X. Sturdivant, *Applied Logistic Regression* (pp. 145-150). Wiley-Blackwell.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Methods to Examine the Scale of a Continuous Covariate in the Logit. In *Applied Logistic Regression* (pp. 94-107). Wiley-Blackwell.

King, J. E. (2008). Binary Logistic Regression. In J. W. Osborne, *Best practices in quantitative methods* (1st ed.). Sage Publications.

King, J. E. (2008). Collinearity. In J. W. Osborne, *Best practices in quantitative methods* (1st ed., pp. 379-380). Sage Publications.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). Collinearity. In *Applied regression analysis and other multivariable methods* (pp. 358-372). Cengage Learning.

Knuth, D. E. (1998). *The Art of Computer Programming Volume 2: Seminumerical Problems* (Vol. 2). Addison-Wesley.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Logistic Regression, Poisson Regression, and Generalized Linear Models. In *Applied Linear Statistical Models* (pp. 555-640). McGraw-Hill.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Multicollinearity and Its Effects. In *Applied Linear Statistical Models* (pp. 278-289). New York: McGraw-Hill.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Multicollinearity Diagnostics - Variance Inflation Factor. In *Applied Linear Statistical Models* (pp. 406-410). New York: McGraw-Hill.

L'Ecuyer, P., & Panneton, F. (2009). F2-Linear Random Number Generators. In C. Alexopoulos, D. Goldsman, & J. R. Wilson (Eds.), *Advancing the Frontiers of Simulation* (pp. 169–193). Springer. doi:https://doi.org/10.1007/b110059_9

L'Ecuyer, P., & Simard, R. (2007). TestU01: A C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software, 33*(4).

L'Ecuyer, P. (2012). Random Number Generation. In J. E. Gentle, W. K. Hardle, & Y. Mori (Eds.), *Handbook of Computational Statistics* (pp. 35-71). Springer.

Lieberman, M., & Morris, J. (2014). The Precise Effect of Multicollinearity on Classification Prediction. *Multiple Linear Regression Viewpoints*, 5-10.

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Transactions on Modeling and Computer Simulation*, 3-30.

Midi, H. S., & Rana, S. (2013, May 28). Collinearity diagnostics of binary logistic regression model. *Laboratory of Applied and Computational Statistics*, 253-267. Retrieved 2021, from Taylor & Francis Online: https://www.researchgate.net/profile/Saroje-Sarkar/publication/261667769_Collinearity_diagnostics_of_binary_logistic_regression_model/links/56472c7b08ae54697fbb9c62/Collinearity-diagnostics-of-binary-logistic-regression-model.pdf

Peduzzi, P., Concato, J., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 1373-1379.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression. *The Journal of Educational Research*, 3-14. Retrieved 2021, from https://www.researchgate.net/profile/Joanne-Peng-4/publication/242579096_An_Introduction_to_Logistic_Regression_Analysis_and_Reporting/links/0deec5374c228b7fa1000000/An-Introduction-to-Logistic-Regression-Analysis-and-Reporting.pdf

Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics*, 705-724.

Saito, M., & Matsumoto, M. (2008). SIMD-Oriented Fast Mersenne Twister: a 128-bit Pseudorandom Number Generator. In A. Keller, S. Heinrich, & H. Niederreiter (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2006.* Springer.

Schisterman, E. F., Perkins, N. J., Mumford, S. L., Ahrens, K. A., & Mitchell, E. M. (2017, January). *Collinearity and causal diagrams – a lesson on the importance of model specification.* Retrieved October 12, 2021, from National Center for Biotechnology Information, U.S. National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5131787/

Senaviratna, N. A., & Cooray, T. M. (2019, October 1). *Diagnosing Multicollinearity of Logistic Regression Model.* Retrieved from Asian Journal of Probability and Statistics: https://www.journalajpas.com/index.php/AJPAS/article/view/30132

Senaviratna, N. A., & Cooray, T. M. (2021, February 27). *Multicollinearity in Binary Logistic Regression Model.* Retrieved from Theory and Practice of Mathematics and Computer Science Vol. 8: https://stm.bookpi.org/TPMCS-V8/article/view/422

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, 1099-1104. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1553-2712.2011.01185.x

Tabachnick, B. G., & Fidell, L. S. (2013). Classification of Cases. In B. G. Tabachnick, & L. S. Fidell, *Using multivariate statistics* (6th ed., pp. 469-470). United States of America: Pearson.

Tabachnick, B. G., & Fidell, L. S. (2013). Correlation. In B. G. Tabachnick, & L. S. Fidell, *Using Multivariate Statistics* (p. 56). Pearson.

Tabachnick, B. G., & Fidell, L. S. (2013). Logistic regression. In *Using multivariate statistics* (pp. 439-509). Pearson.

Tabachnick, B. G., & Fidell, L. S. (2013). Multicollinearity and Singularity. In 6th (Ed.), *Using Multivariate Statistics* (pp. 88-91). United States of America: Pearson.

Tabachnick, B. G., & Fidell, L. S. (2013). Practical Issues. In *Using multivariate statistics* (6th ed., pp. 444-446). United States of America: Pearson.

Tabachnick, B. G., & Fidell, L. S. (2013). Review of Univariate and Bivariate Statistics. In B. G. Tabachnick, & L. S. Fidell, *Using Multivariate Statistics* (pp. 33-37). Pearson.

Vasu, E. S., & Elmore, P. B. (1975, January 23). *The Effect of Multicollinearity and the Violation of the Assumption of Normality on the Testing of Hypotheses in Regression Analysis.* Washington, D.C.: US DEPARTMENTOF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION. Retrieved from ERIC: https://eric.ed.gov/?id=ED106341

Zeng, G., & Zeng, E. (2019, March 28). *On the relationship between multicollinearity and separation in logistic regression.* Retrieved from Taylor & Francis Online: https://www.tandfonline.com/doi/abs/10.1080/03610918.2019.1589511?journalCode=lssp20#:~:text=Multicollinearity%20and%20separation%20are%20two%20major%20issues%20in%20logistic%20regression.&text=We%20analytically%20prove%20that%20multicollinearity%20means%

BIBLIOGRAPHY

Aaron Wayne Ghaner

Candidate for the Degree of

Master of Science Mathematics

Thesis: REPRECUSSION OF MULTICOLLINEARITY IN BINARY LOGISTIC REGRESSION

Major Field: Mathematics

Education: Bachelor of Science Traditional Physics

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio in 2022.

ADVISER'S APPROVAL: Dr. Douglas Darbro

7/27/2022