2-26-2021 2:30 PM

Sample Size Formulas for Estimating Risk Ratios with the Modified Poisson Model for Binary Outcomes

Zhenni Xue, *The University of Western Ontario*

Supervisor: Zou, Guangyong, *The University of Western Ontario*
Co-Supervisor: Choi, Yun-hee, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Sample Size Formulas for Estimating Risk Ratios with the Modified Poisson Model for Binary Outcomes

Zhenni Xue

Supervisor: Zou, Guangyong, *The University of Western Ontario*
: Choi, Yun-hee, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

Sample size estimation is usually the first step in planning a research study. Too small a study cannot adequately address the objectives, while too large a study may waste resources or be unethical. For binary outcomes, several sample size estimation methods are available based on logistic regression models, which focusing on odds ratios. In prospective studies, risk ratios are preferable for ease of interpretation and communication. In this thesis, we compared the power difference between the logistic regression model and the modified Poisson regression model via simulation studies. We then proposed sample size estimation formulas based on the modified Poisson regression model for estimating risk ratios. Simulation results suggested that both models have similar performance in terms of Type I error and power. The empirical evaluation indicated that the proposed sample size formulas are reliable in a wide range of scenarios. The sample size estimation procedure was illustrated using a subset of data from the Diabetes Control and Complications Trial.

Keywords: binary data; sample size; risk ratio; odds ratio; logistic regression; power; study design

Summary for the Lay Audience

Medical and epidemiological research rests largely on assessment of risks. One key measure in such studies is the ratio of odds, which is commonly estimated using logistic regression models. However, ratio of odds has been commonly interpreted as ratio of risks. Since by definition odds is larger than risk numerically, this practice can exaggerate study results, especially when risk of event is not rare as in many prospectively studies. The modified Poisson regression model was proposed as a method to estimate risk ratio directly. The model has become increasingly applied in medical and epidemiological research. To facilitate its use, this thesis compares power of the modified Passion regression to that of the logistic regression using simulation studies. The results suggest equivalent power between the two models. The thesis further proposed and evaluated sample size formulas based on the modified Poisson model. Simulation results suggest the formulas performed well, providing an important tool for study planning.

Acknowledgements

I would like to express thanks to my supervisors, Professor Guangyong Zou and Professor Yun-hee Choi, of the Department of Epidemiology and Biostatistics at the University of Western Ontario for their scientific guidance and encouragement. This thesis would not have been possible without their insights and professional comments.

I thank Grace Lu and Bingyu Nie for mental support.

I thank my family for financial support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

RD	Risk difference
RR	Risk ratio
OR	Odds ratio
NP	Nominal power
EP	Empirical power
SE	Standard error
CI	Confidence interval
mpoisson	The modified Poisson model
logistic	The logistic regression

Chapter 1 Introduction

Any research inquiry begins with study planning. To ensure scientific validity, a study should be designed to meet clearly defined objectives. Determination of sample size is an important component in the design of epidemiological and medical studies. A study should be large enough to address the research questions but not too large to be wasteful or unethical in putting participants in harm.

Effect measure is an important element of study planning and sample size estimation. The choice of appropriate effect measures is crucial for research to address meaningful objectives adequately. This decision usually rests on the types of outcome data. Binary outcomes are prevalent in epidemiological and medical inquiries. Thus, this thesis focuses on this outcome. There are various effect measures for binary outcomes, including odds ratio, risk difference, and risk ratio. Due to its availability in case-control studies and its connection with logistic regression models, odds ratios have been the predominant effect measure when the outcome is binary. However, the risk difference and risk ratio are easier to understand and communicate. Regression models for risk differences and risk ratios are less well known. The modified Poisson regression model for risk ratios has been increasingly adopted in epidemiological and medical studies (Spiegelman & Hertzmark, 2005; Zou, 2004).

A variety of sample size formulas exist for the logistic regression models with the odds ratio as the parameter of interest. Whittemore (1981) proposed a sample size formula by applying the maximum likelihood procedure for the logistic regression. Hsieh et al. (1998) described sample size methods based on the odds ratio for two-group comparison studies. There is a paucity of sample size formulas based on the modified Poisson model for estimating risk ratios.

This chapter begins with the description of the binary outcome, followed by a discussion on the choice of effect measures for binary outcomes in Section 1.2. Section 1.3 summarizes the effect measures related to regression models. Section 1.4 provides a brief review of the literature concerning sample size estimation. The final section describes the objectives and organization of the thesis.

1.1. Binary outcome in medical and epidemiological research

Binary data can arise in at least two ways. First, binary outcomes arrive naturally to describe two states of nature. Examples include the diagnostic test for a subject being positive or negative, presence or absence of a disease condition, and alive or dead of a subject. Second, binary outcomes can arise from dichotomizing continuous data. The use of dichotomized data has several drawbacks, including loss of information, subsumption of variability for the original outcome, and concealment of variable associations (Altman & Royston, 2006). Thus, dichotomization requires adequate justification, often based on well-accepted criteria and for ease of interpretation. For instance, blood pressure can be categorized into hypertensive if a blood pressure above 130/90 mmHg, or normotensive otherwise. Another example is that diabetes is defined by a glucose level greater than 125 mg/dL.

In practice, the event of interest is usually denoted as 1 while the non-event is denoted as 0 in data analysis. This designation should not be used lightly, as it could have important implications for interpreting research results. Sheps (1958) discussed that the interpretation of results could be affected by the choice of the reference state. For example, one could report a small relative difference in survival rate but a large relative difference in death rate for the same male and female groups. While the absolute difference of the survival or death rate between males and females is the same for death and survival, the denominator state of relative comparison changed the interpretation. Treating alive or dead as the state of interest brings a vastly different impression, although the two states are complementary.

1.2. Effect measures for binary outcomes

The general goal of a study is to assess the associations between exposures and outcomes. In this thesis, we focus on situations where both exposure and outcome are binary, but covariates can be of multiple types. Measures of these associations are commonly referred to as effect measures. For ease of communications, effect measures for binary outcomes are usually defined in the context of binary exposure, such as exposure versus unexposed, treatment or control, and one level versus another level of continuous exposure. The common effect measures include the risk difference (RD), risk ratio (RR), and odds ratio (OR). Risk difference is defined as the absolute difference in probabilities of outcomes between two exposure groups. Risk ratio is defined as the ratio of the two probabilities, while OR is the ratio of the two odds, with odds defined as a probability divided by its complement. In the case of rare events of event probability less than 0.1, RR is approximately equivalent to OR in magnitude.

Regarding the effect measure selection, Lachin (2011, p. 21) pointed out that all three types of effect measures could reflect differences between groups, and the measurement choice should not influence research conclusions when the sample size is large. Nonetheless, there is empirical evidence that the choice of effect measures could impact the results of a clinical trial. For example, Bobbio et al. (1994) reported on the willingness of physicians prescribing drugs and showed that the decision of prescription depended on the choice of effect measures. Five types of effect measures were presented to physicians. For each measurement, physicians rated their willingness to prescribe a drug on a 0 to 100 scale, and more than half of the physicians would tend to prescribe based on the RR. The relative comparison of risks of getting diseases leaves the impression of a more significant benefit if the medicine is used to treat patients compared to other measures. Bobbio et al. (1994) also reported that physicians might misinterpret the reported results, as many of them were not trained to differentiate the differences among effect measures.

Walter (2000) categorized the properties of the effect measures into six aspects: simplicity, symmetry, range of predicted event rates, biased or unbiased estimate, estimation efficiency, and estimation model availability. Each effect measure has favorable features. For instance, the RD is simple and easy to interpret. In contrast, the interpretation of OR can be confusing for non-statisticians, but the OR is applicable in various clinical studies due to its mathematical properties. Walter (2000) suggested that the choice of effect measure should not rely solely on the mathematical convenience or the anticipation of the comprehensibility of a specific effect measure when selecting an effect measure to use. A suitable effect measure should be chosen based on the fitness of the actual data to a specific effect measure model.

1.3. Estimation of effect measures with multivariable regression models

The difficulty of estimating effect measures often arises in studies involving multiple independent variables in which effect measures of interest are usually estimated using regression models. Logistic regression is widely accepted and used to estimate OR in prospective, retrospective, and case-control studies. The logistic regression model connects the probability of binary outcome with a linear combination of the predictors using a logit function, where the logarithm of ORs is estimated with the maximum likelihood method.

When risk ratios are of interest in prospective studies, the RR can be estimated using the log-binomial model (Wacholder, 1986) or a Poisson model (McNutt et al., 2003). However, the former can encounter convergence problems during maximum likelihood iteration, while the latter results in overestimated standard errors. The modified Poisson model with robust error variance overcomes both problems (Zou, 2004). This model has also been extended to studies with correlated binary outcomes (Yelland et al., 2011; Zou & Donner, 2013).

Inspired by the modified Poisson model for risk ratios, Cheung (2007) proposed the modified least-squares regression for risk differences using the robust standard error in the binomial regression with an identity link. Spiegelman and Hertzmark (2005) suggested using estimates from the Poisson regression and log-binominal model as the starting values for the iteration algorithm to improve the efficiency of both the modified Poisson model and the modified least square model.

1.4. Estimation of effect measures and sample size

There are various approaches to determine the sample size for different effect measures. For example, Whittemore (1981) found a sample size for the OR by approximating the Fisher information matrix in logistic regression with a small response probability. Hsieh et al. (1998) presented sample size formulas for the OR for comparing two groups using the logistic regression model. Donner (1984) reviewed sample size formulas for assessing risk differences and risk ratios in randomized control trials. There is a paucity of sample size formulas for RD and RR in the context of regression models.

Four parameters are commonly included in the power analysis: effect size, significance level, statistical power, and sample size. The significance level is the probability of rejecting the null hypothesis when it is true, and is also called the alpha level (α), which is commonly set at 0.05. The value of any of the remaining three parameters can be determined by knowing the other two.

1.5. Objectives and outlines

The primary focus of this thesis is on power and sample size with respect to the modified Poisson regression model. We aim to derive a sample size formula for estimating RR using the modified Poisson model. The performance of the derived sample size is assessed through simulations and illustrated using real data applications.

The thesis has three main objectives:

1. To assess the power of the modified Poisson model in comparison with the logistic regression model.
2. To derive a simple and closed-form sample size formula for estimating risk ratios using the modified Poisson model.
3. To assess the proposed sample size formula using simulation studies. The assessment environment is assumed to fit the modified Poisson model.

The thesis has six chapters. Chapter 2 discusses the three major effect measures and concentrates on the benefits and drawbacks of three regression models for binary outcomes, the general principle of sample size, and various sample size equations. Chapter 3 contains the development of the sample size for estimating RR. In Chapter 4, we conduct two simulation experiments for power comparisons. In Chapter 5, an application is designed to test the feasibility of the sample size and power formulas derived from Chapter 3, using a subset of data from the Diabetes Control and Complications Trial (Diabetes Control and Complications Trial Research Group, 1993). In Chapter 6, we discuss the implications and limitations of the research as well as potential future studies.

Chapter 2 Literature Review

This chapter reviews the literature, beginning with the three common effect measures in Section 2.1. Section 2.2 discusses the regression models which are used for estimating effect measures. The chapter closes with sample size formulas based on two-sample comparison and maximum likelihood procedure in Section 2.3.

2.1. Effect measures for binary outcomes

Recall that there are three common effect measures with binary outcomes: the RD, the RR, and the OR. Table 2.1 presents a 2 x 2 frequency table from a total number of subjects of n. Let x serve as an indicator of exposure (1) or non-exposure (0); y is the possible binary outcome of an individual, with 1 denoting event and 0 denoting non-event.

Table 2.1 General 2 x 2 table for binary x and y

	$y = 1$	$y = 0$
$x = 1$	a	b
$x = 0$	c	d

The three measures can be defined as follows:

$$RD = a/(a+b) - c/(c+d) = P_1 - P_0,$$

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{P_1}{P_0},$$

$$OR = \frac{a/b}{c/d} = \frac{P_1/(1-P_1)}{P_0/(1-P_0)},$$

where a, b, c, d are the numbers of individuals for (x, y) combinations, and P_1 and P_0 are probabilities of $y = 1$ for the exposed and non-exposed x groups, respectively.

Each effect measure has its own properties. The RD is an absolute and straightforward measure, quantifying the net risk (Sinclair & Bracken, 1994). The estimator of the RD is unbiased if the two variables are independent binomials. In contrast to the RD, the estimators for RR and OR are not unbiased, but they are consistent, meaning that the bias is negligible with a large sample size. The RD has a symmetric property meaning that the group difference stays unchanged or has a sign difference when interchanging successes and failures. For example, if the disease probabilities of the treatment and control groups are 0.1 and 0.2, respectively, the RD is -0.1. Switching the two groups in the RD calculation gives RD = 0.1, which only has a negative sign difference with RD = -0.1. The RD can produce risk probabilities that are out of range. If RD is 0.1 when using the treatment group as the baseline, and the mortality of the control group is 0.09, the mortality of the treatment group is -0.01.

The risk ratio is an effect measure commonly used in prospective studies, especially in randomized clinical trials. Compared to OR, clinicians would prefer to use the RR when comparing the risk of disease between the treatment and control groups instead of asking for the odds comparison (Sinclair & Bracken, 1994). Another benefit is that RR is straightforward in interpretation and easy to explain by the public.

As pointed out by Greenland (1987), the risk ratio is the ratio of two cumulative incidences between exposed and unexposed groups, and it is only interpretable as the effect on average risk. The RD can be interpreted as an average effect on risk or an effect on average risk, and the OR can represent neither. The RR produced a predicted event probability out of range. If an individual patient in the control group has an outcome event probability of 0.5, the patient will have an outcome event probability of 1.5 in the treatment group when RR equals 3 (treatment vs. control). In addition, the RR is not symmetric, as the RR of the mortality rate between the two groups is not a reciprocal of the RR of the survival rate between the same groups.

The odds ratio has some advantages that the other two measures do not have. Unlike the RR, the OR is symmetrical when interchanging the two groups (Walter, 2000). Switching two groups brings an OR that is the reciprocal of the pre-interchanged OR. Thus, when the log function is applied to the OR, researchers only need to change the sign of the log(OR). Predictions of probability based on OR has a restriction of [0,1], while that based on RR and RD may be out of the 0 to 1 range.

The odds ratio is the most popular measure among the three measures in statistical analyses. One of the reasons is that researchers can estimate the OR in prospective, retrospective, and case-control studies (Walter, 2000). The ratio between odds of exposure in cases and controls and the ratio between odds of the outcome in exposed and unexposed groups are equal mathematically (Cornfield, 1951). If the defined study populations are the same, researchers could have the same OR estimates from a case-control study or a prospective study for exposure and disease. Cornfield (1951) pointed out that the OR could be used for estimating the RR when the disease is rare, but the disease and control groups should represent the corresponding groups in the general population. Another reason for the popularity of OR is that the OR can be estimated and tested using the widely adapted logistic regression.

The odds ratio may have some disadvantages in practice. The OR is the indirect measure for estimating risks compared to the other two effect measures. Greenland (1987) argued that the OR is a useful effect measure when applying it in estimating the RR, as only the RD or RR can directly measure the influence of an intervention on average risk. However, the OR estimates can be influenced by the outcome event probability of the control group, so using the OR as a substitute of RR can be misleading in communicating the results from cohort studies (Nurminen, 1995; Sinclair & Bracken, 1994). Another downside of the odds ratio is its non-collapsible property. The non-collapsibility refers to when the marginal and conditional odds ratios are different in magnitudes even in the absence of confounding (Greenland et al., 1999).

It is widely recognized that when the outcome event is rare, the OR can approximate the RR. However, under the common disease or the unstable probability of exposure environment, the approximation can be biased (Nurminen, 1995). The OR approximately equals the RR when the baseline risk is less than 10% (Sinclair & Bracken, 1994). However, there are other requirements for the approximation besides the low outcome event probability. Greenland (1987) discussed that the approximation requires the probability of $y = 1$ to be small in each covariate category, not only for the overall probability to be small. This relationship can be illustrated as following:

$$\frac{a/b}{c/d} = \left(\frac{a/(a+b)}{c/(c+d)}\right)\left(\frac{1+a/b}{1+c/d}\right).$$

The OR can be written as the RR times an additional term. If the odds from each group do not exceed the value M, the value of the additional term is within the $(1 \pm \mathrm{M})$ range. The approximated RR is closer to the OR with a smaller M value.

In practice, it is not uncommon to find examples where the OR has been interpreted as RR. For example, in a study of the effects of race and sex on the referral rate for cardiac catheterization, an estimate of $\mathrm{OR} = 0.60$ was interpreted as RR, suggesting a black female patient would have 40% less probability of being referral compared to a white counterpart (Schwartz et al., 1999). This result caused heated debate regarding racial discrimination in several US mass media, while the RR estimate was 0.93 and the referral rate reduction was only 7%.

2.2. Multiple regression models for binary outcomes

The effect measures in the studies with a binary outcome can be estimated by using regression models. This is important for adjusting for potential confounding and/or improve estimation precision. We review the logistic regression, log-binomial model, and the modified Poisson model. The logistic regression is widely used in analyses to estimate the OR. The log-binomial and modified Poisson models are for the RR, which is the effect measure that we focus on in this thesis.

2.2.1. Logistic regression for estimating odds ratios

Consider a regression with k covariates. Let $x_1, \ldots x_k$ be the individual covariates. The logistic regression is usually written as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{2.1}$$

where p is the probability of outcome events given covariates. The regression coefficient β represents the log of odds or log of odds ratio in the logistic regression.

The logistic regression has several assumptions, and some of the assumptions are similar to those of linear regression. Linear regression requires a linear relationship between dependent and independent variables. The logistic regression does not require a linear relationship between the dependent and independent variables. Instead, the linear relationship is assumed between log odds and independent variables. Another assumption for both linear and logistic regressions is that the observations should be independent of each other. The last assumption is the absence of multicollinearity for both regression models. Multicollinearity can be detected by using the variance inflation factor (VIF). Mansfield and Helms (1982) recommended that the VIF should not be too far from 1.0. The common cutoff for determining multicollinearity is VIF=10 (Hair, et al., 1998, p. 200).

Vinttinghoff et al. (2012, pp. 141-144) pointed out several advantages of using the logistic regression to estimate the OR. First, the predicted outcome event probability from the logistic regression will not be out of [0,1] range. Another advantage of logistic regression is that the coefficients can be expressed as the log of the odds or log of the OR. The model has a multiplicative property so that the odds for a treatment group could be calculated using the anti-log function of the corresponding regression coefficient and the odds of the control group. The popularity of the logistic regression is due largely to the software availability of model fitting.

The logistic regression model has been regarded as a universal method in different epidemiological studies. However, the model has potential disadvantages. Sinclair and Bracken (1994) considered that the OR could be misinterpreted or even mislabeled as RR. Misinterpreting may lead to the confusion of the size of the covariate effect. As evidenced in the example from Sinclair and Bracken (1994), a study of the association between hemoglobin level and mortality described a relative comparison of mortality between low and high hemoglobin groups, but the actual effect measure used was OR. The calculated true RR from the same study was about half of the reported number. Again, misinterpretating OR as RR has been vividly demonstrated by the high-profile study regarding influences of race and sex on the referral rate for cardiac catheterization (Schwartz et al., 1999).

The non-collapsibility of the OR, as described in Section 2.1, is another problem. Having the stratification variable *z* be independent of other covariates cannot guarantee the collapsibility when the link function is logistic (Greenland et al., 1999). Falsely interpreting the marginal effect as a stratum-specific effect could lead to an inappropriate conclusion.

The consequence of non-collapsibility was emphasized by Gail et al. (1984) and Neuhaus and Jewell (1993) that the estimation of the regression coefficients would be affected by the omitted covariate in randomized studies. Neuhaus and Jewell (1993) demonstrated that the bias can arise even when the omitted covariate is independent of the other covariates. The direction of the bias relates to the concave or convex status of the link functions. The magnitude of the bias depends on the variance of the omitted covariate effect. The larger the variability, the larger the bias. Omitting covariates that are correlated to those included covariates also affects the regression coefficients, as the included and omitted covariates may not be conditionally independent given the outcome. Researchers choose covariates based on their needs and they may not include all possible variables related to the outcome, and the regression coefficient estimation can be biased.

2.2.2. Log-binomial model for estimating risk ratios

It is not hard to find many studies with common outcome. In such cases, using OR to approximate RR can be misleading. The RR could be directly estimated using log-binomial models. Such a model uses a log link to connect covariates and the probability of the outcome events, whereas the logistic regression applies a logit link in between. The model can be written as follows (Wacholder, 1986):

$$\log(p) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k, \tag{2.2}$$

where β_0 represents the log of baseline risk and $\beta_1, \ldots, \beta_k$ are the log of the risk ratios.

As a member in the family of generalized linear models, one advantage of the log-binominal model is that it uses maximum likelihood estimation approach and thus provides efficient estimates of risk ratios. The log-binominal model has a good property that excluding or including uncorrelated independent variables to the variable of interest would not materially change the estimate of risk ratio for the variable of the interest, due to the log link function (Neuhaus & Jewell, 1993).

However, the log-binomial model may face the issue of non-convergence. The convergence error during the maximum likelihood iteration could occur when the right side of Equation (2.2) becomes higher than zero for some individual observations. The RRs are not estimable from the log-binomial model in this case. The problem implies that the selected model is inappropriate for fitting the data (Wacholder, 1986).

2.2.3. Modified Poisson model for estimating risk ratios

McNutt et al. (2003) proposed estimating risk ratio by the Poisson regression, and found that this model overestimates standard errors, due to the misspecification of the Poisson model for binary outcome. The overestimation happens when the disease is common. The standard errors from the Poisson model and the log-binomial model are similar when the disease is rare, as $p \approx p(1-p)$. To correct for model misspecification, Zou (2004) proposed the modified Poisson model using a robust sandwich estimator for variance estimation.

The robust sandwich variance estimator can be expressed as $I^{-1}JI^{-1}$, where I is the Fisher information matrix from the Poisson regression, and J is the empirical estimate of the covariance matrix. For the data in Table 2.1, we can fit the modified Poisson model with only one binary covariate x:

$$\log\big(E(Y)\big) = \beta_0 + \beta_1 x,$$

where the β_0 is the intercept, and β_1 is the regression coefficient of the interest. The binary outcome vector Y and the design matrix X are

$$Y = \begin{bmatrix} 1_a \\ 0_b \\ 1_c \\ 0_d \end{bmatrix} \qquad X = \begin{bmatrix} 1_a & 1_a \\ 1_b & 1_b \\ 1_c & 0_c \\ 1_d & 0_d \end{bmatrix},$$

where the 1_a is a vector of ones with the length of a, and the 0_b is a vector of zeros with the length of b. The Fisher information matrix I can be showed as

$$I = \begin{bmatrix} (a+b)e^{\beta_0+\beta_1} + (c+d)e^{\beta_0} & (a+b)e^{\beta_0+\beta_1} \\ (a+b)e^{\beta_0+\beta_1} & (a+b)e^{\beta_0+\beta_1} \end{bmatrix},$$

where $e^{\beta_0+\beta_1}$ and e^{β_0} are estimated as $a/(a+b)$ and $c/(c+d)$, respectively. The inverse of the Fisher information matrix can be estimated as

$$I^{-1} = \begin{bmatrix} \frac{1}{c} & -\frac{1}{c} \\ -\frac{1}{c} & \frac{1}{c}+\frac{1}{a} \end{bmatrix}.$$

The covariance matrix, J, can be empirically estimated as

$$J = X'diag(residuals)^2X$$

$$= \begin{bmatrix} a\left(1-\frac{a}{a+b}\right)^2 + b\frac{a^2}{(a+b)^2} + c\left(1-\frac{c}{c+d}\right)^2 + d\frac{c^2}{(c+d)^2} & a\left(1-\frac{a}{a+b}\right)^2 + b\frac{a^2}{(a+b)^2} \\ a\left(1-\frac{a}{a+b}\right)^2 + b\frac{a^2}{(a+b)^2} & a\left(1-\frac{a}{a+b}\right)^2 + b\frac{a^2}{(a+b)^2} \end{bmatrix},$$

where the residuals are obtained as $Y-\mu$, and μ is estimated by $\mathrm{e}^{\beta_0+\beta_1 x}$. The diag(residuals) in the covariance matrix represents the diagonal matrix of residuals. Putting the matrices together, the robust sandwich variance estimator is estimated as

$$I^{-1}JI^{-1} = \begin{bmatrix} \frac{1}{c} & -\frac{1}{c} \\ -\frac{1}{c} & \frac{1}{c}+\frac{1}{a} \end{bmatrix} \begin{bmatrix} \frac{ab}{a+b}+\frac{cd}{c+d} & \frac{ab}{a+b} \\ \frac{ab}{a+b} & \frac{ab}{a+b} \end{bmatrix} \begin{bmatrix} \frac{1}{c} & -\frac{1}{c} \\ -\frac{1}{c} & \frac{1}{c}+\frac{1}{a} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d}{c(c+d)} & -\frac{d}{c(c+d)} \\ -\frac{d}{c(c+d)} & \frac{b}{a(a+b)}+\frac{d}{c(c+d)} \end{bmatrix}.$$

Then, the (2,2)th element of the $I^{-1}JI^{-1}$ brings the estimated variance of the regression coefficient of the interest:

$$\widehat{\mathrm{Var}}\left(\widehat{\beta_1}\right) = \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.$$

The robust sandwich estimator provides consistent covariance estimates when heteroskedasticity exists, and it can correct the inconsistency due to model misspecification (White, 1980). The modified Poisson regression was shown to provide equivalent variance estimates with that of the log-binominal regression, but without convergence problems.

Ritz and Spiegelman (2004) suggested to use the robust sandwich estimator in a marginal model to adjust a random effect from omitting a covariate when the response probability follows a Poisson distribution. The modified Poisson model uses a working marginal Poisson distribution (Zou & Donner, 2013), and the estimation of the standard errors is not influenced by omitting a covariate when confounding is absent as in randomization trials. Ritz and Spiegelman (2004) also recommended a generalized estimating equation (GEE) approach in the marginal model when the link function is the identity or log, as it provides consistent regression coefficient estimates when the working correlation structure is inappropriate.

Yelland et al. (2011) evaluated the modified Poisson model with clustered prospective data using simulation. In the study, the exposure status was independent of the clusters, and it was assigned to the individual level. The GEE procedure with an exchangeable correlation structure was applied to accommodate the clustering. The performance of the modified Poisson regression with the GEE was similar to that of the log-binomial model when the log-binomial model did not have a convergence problem.

Zou and Donner (2013) extended the modified Poisson model to the correlated binary outcomes for longitudinal or clustered randomized trial studies, where the entire clusters were randomized into either exposed or non-exposed group. The sandwich variance estimator with the adjustment based on clusters was applied, and the GEE was also used. The model performed well with the correlated binary outcome. Meanwhile, the model could accommodate situations in which the log-binomial regression had convergence problems.

The modified Poisson model has some other benefits. It is appealing to many clinicians because it estimates the RR, which is easier to interpret. As discussed in Section 2.1, the RR is collapsible, so that the modified Poisson model could benefit from the collapsibility. The modified Poisson model can still be applied when the log-binomial model faces a convergence issue, as the modified Poisson model is not likely to have convergence difficulties (Yelland et al., 2011).

Zou and Donner (2013) suggested that the modified Poisson model may not be capable of predicting individual risks, because for an individual i ($i = 1 \ \ldots n$), the right side of the equation $\log(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$ could be estimated greater than 0, which results in meaningless individual probability $p_i > 1$.

Compared to the log-binomial model, the modified Poisson model rarely has convergence problems and can be used as an alternative model to the log-binomial model in various situations to estimate the RR.

2.3. Sample size and power for studies with binary outcomes

The sample size determination is an essential element in planning a medical research study. The literature has suggested many methods in finding adequate sample sizes. Based on odds ratios, Whittemore (1981), Lachin (1981), and many others have addressed the sample size formula with the two-group comparison and maximum likelihood methods. We review the general theory and various methodologies in this section.

2.3.1. General principle

Consider a hypothesis testing to detect the difference between two means. Let μ represent the difference of two group means; and the statistic S be a consistent estimator of μ. The null hypothesis is μ equals to μ_0, and the statistic follows $N(\mu_0, \Sigma_0^2)$ under the null. Under the alternative hypothesis, μ equals to μ_1, and it is distributed with $N(\mu_1, \Sigma_1^2)$. The Σ_0^2 and Σ_1^2 can be written as σ_0^2/n and σ_1^2/n, respectively. The σ_0^2 and σ_1^2 are the variance of the individual observations under the null and the alternative, respectively, and n stands for the total sample size. Two types of errors exist in hypothesis testing. The α represents the Type I error or significance level, which is the probability of rejecting the null hypothesis when it is true. The Type II error is defined as the probability of not rejecting the null hypothesis when the alternative hypothesis is true. The γ represents the power which is related to the Type II error ($\gamma = 1 - \text{Type II error}$), and the power is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true. In hypothesis testing, a test statistic is used to evaluate the probability that the observations could happen by chance (Lachin, 2011, p.87). For rejecting H_0, the two-tailed significance test for μ needs to satisfy the following relationship under the null hypothesis (Lachin, 2011, pp. 87-91)

$$|T| = \left|\frac{\mathrm{S} - \mu_0}{\sigma_0}\right| > \mathrm{Z}_{1-\alpha/2}.$$

From the test, the power (γ) can be expressed as follows:

$$\gamma = \Pr\left[|T| > Z_{1-\alpha/2}\middle|\mathrm{H}_1\right], \tag{2.3}$$

where T represents the test statistic that follows a standard normal distribution under the null and $Z_{1-\alpha/2}$ is the critical value of the standard normal distribution at α significance level (Lachin, 2011, p. 89).

For finding a sample size, the power of a hypothesis test can be used. The calculation of the test statistics T contains the variance of the estimated S. The estimation of the variance involves a sample size n. By inverting Equation (2.3), n can be found. Investigators can use various hypothesis tests, such as the Chi-square test, Wald test, and Score test, but the general principle for finding the sample size remains the same.

2.3.2. Two-group comparison studies

Many clinical trial studies involve two-group comparisons, and the hypothesis test is based on the difference between the two group means. Lachin (1981; 2011) derived the formulas for the sample size *n* and the power γ from the two-tailed test by solving for n and Z_γ from the difference $|\mu_1 - \mu_0| = Z_{1-\alpha/2}\Sigma_0 + Z_\gamma\Sigma_1$. as the following:

$$n = \left(\frac{Z_{1-\alpha/2}\sigma_0 + Z_\gamma\sigma_1}{\mu_1 - \mu_0}\right)^2, \tag{2.4}$$

$$Z_\gamma = \frac{\sqrt{n}|\mu_1 - \mu_0| - Z_{1-\alpha/2}\sigma_0}{\sigma_1}.$$

Lachin (1981) also extended the general sample size formula to test the difference between two proportions. The null hypothesis is that there is no difference between the two proportions. The alternative hypothesis is that the two proportions are different. The variances under the null and alternative hypotheses are calculated to replace σ_0^2 and σ_1^2 in Equation (2.4).

For the risk ratio, Donner (1984) reviewed sample size formulas with the randomized clinical trial design. The null hypothesis is $\mathrm{H}_0: RR = 1$. The test statistics come from the Chi-square test. The square root of the Chi-square statistics follows a standard normal distribution. The sample size for RR is obtained by

$$n = \left(\frac{Z_{1-\alpha/2}\sqrt{P_0(1+RR)(1-\bar{P})} + Z_\gamma\sqrt{P_0\left(1+RR-P_0(1+RR^2)\right)}}{P_0(1-RR)}\right)^2, \quad (2.5)$$

where $P_0, P_1, \bar{P}$ are the probability of the outcome event for the control group, the intervention group, and the average of the outcome probability of the two groups, respectively.

Hsieh et al. (1998) presented a sample size formula based on the logistic regression for OR with one covariate, assuming the covariate follows a standard normal or binomial distribution. The covariate distributions in each binary response group and the overall covariate distribution are presumed to be the same. When the covariate is binary, the sample size for a simple logistic regression can be written in terms of P_0 and P_1, where P_0 and P_1 are also the outcome event probabilities from the control and intervention groups, respectively. The P_x is the probability for $x = 1$, and p is the overall prevalence of the outcome event. With a two-tailed test, the sample size is

$$n = \frac{\left(Z_{1-\alpha/2}\sqrt{p(1-p)/P_x} + Z_\gamma\sqrt{P_0(1-P_0) + P_1(1-P_1)(1-P_x)/P_x}\right)^2}{(P_0-P_1)^2(1-P_x)}.$$

To account for the situation in which multiple risk factors are present, Hsieh et al. (1998; 2003) applied the variance inflation factor (VIF) to the sample size formulas to adjust for more than one covariate situation in the model. The VIF is

$$\mathrm{VIF} = \frac{1}{1-r_{1,2\ldots k}^2}, \quad (2.6)$$

where $r^2_{1,2\ldots k}$ is the coefficient of determination that comes from the model that the factor of interest regresses on the other covariates. The reason for the adjustment is that the null hypothesis is $\mathrm{H_0}: [\beta_1, \beta_2 \ldots \beta_k] = [0, \beta_2 \ldots \beta_k]$, and it is not only $\mathrm{H_0}: \beta_1 = 0$. Hsieh et al. (2003) pointed out that both the variance of the residuals and the variance of other covariates can influence the variance and covariance of the factor of interest, which leads to a power reduction. Thus, the power is adjusted by the VIF to account for the association between the factor of interest and other covariates.

Alam et al. (2010) considered that when the null hypothesis is not true, the assumption by Hsieh et al. (1998) of identical conditional distributions $(X|Y)$ for each response group ($Y = 1$ or $Y = 0$) might not be true. Under the alternative hypothesis, if X follows a normal distribution, the distributions for $(X|Y = 1)$ and $(X|Y = 0)$ can be non-normal, and the variances may vary. In addition to the non-standard two-sample framework, the sample size from Hsieh et al. (1998), which assumes the intercept of the logistic regression is known, is unstable and sensitive to minor changes in the intercept β_0. Alam et al. (2010) also illustrated that the method by Hsieh et al. (1998) is not accurate when the covariate follows a Bernoulli distribution.

2.3.3. Sample size based on maximum likelihood estimation

Apart from the efforts of Hsieh et al. (1998), much of the literature pursued other sample size methods of the odds ratio for the logistic regression model. The variance of the estimator in the sample size approximation can be related to the variance-covariance matrix from the maximum likelihood method. Let θ be a vector of the unknown parameters in the logistic regression model and they are estimated by the maximum likelihood method.

Whittemore (1981) estimated the sample size by approximating the Fisher information matrix of the maximum likelihood estimates in a closed form when the response probability is small and the covariates follow a family of the multivariate distribution. The VIF was applied to reduce the complexity of the variance-covariance matrix of the estimates. Alam et al. (2010) pointed out that the small response probability assumption introduces severe restrictions on regression coefficients. Hsieh (1989) relaxed the condition by simplifying the sample size derived by Whittemore (1981) with the assumption of $[\theta_2 = \cdots = \theta_k = 0]$ and the use of VIF.

Another sample size method was proposed by Self and Mauritsen (1988) based on score statistics under the alternative hypothesis H_1. The standard score statistics, $U(\theta_0)^2/I(\theta_0)$, is evaluated under the null hypothesis $\mathrm{H}_0: \theta = \theta_0$, where θ_0 can be zero or other fixed values. Self and Mauritsen (1988) implemented the standard Taylor series in approximating the score test statistics under the alternative hypothesis and used the non-central Chi-square distribution for approximating the power of the score test.

The score test is one of the likelihood-based tests that can be applied in the sample size estimation. The other two tests are the likelihood ratio test and Wald test. Demidenko (2007) suggested the Wald test should be applied to approximate the sample size because it is commonly used to test the significance of a regression coefficient in the data analysis stage. Demidenko (2007) replaced the variance σ_0^2 under the null hypothesis with the variance σ_1^2 under the alternative hypothesis in the sample size equation based on the Wald test:

$$n = \frac{\left(Z_{1-\alpha/2}\sigma_0 + Z_\gamma \sigma_1\right)^2}{(\beta_k^* - \beta_k^0)^2} = \frac{\left(Z_{1-\alpha/2} + Z_\gamma\right)^2 \sigma_1^2}{(\beta_k^* - \beta_k^0)^2},$$

where β_k^0 is the value of β_k under the null hypothesis, and β_k^* is the value of β_k under the alternative hypothesis. The replacement was proposed because the Wald test statistics uses the variance estimated at MLE, not under the null hypothesis. Using the variance under the null hypothesis may lead to a biased sample size. The variance under the alternative hypothesis is numerically computed depending on the covariate distribution. The sample size method by Demidenko (2007) is beneficial to logistic regression with any types of covariate distributions. However, the sample size for regressions with non-binary covariates requires numerical calculation because there is no closed form.

Unlike the method proposed by Demidenko (2007), Alam et al. (2010) applied variances under both null and alternative hypotheses into sample size determination. This method is a variation of the sample size approach by Whittemore (1981), and the estimated sample size also varies with the changes in the logistic regression intercept. The proposed method provided a better power estimation than the method by Hsieh et al. (1998) under the same settings when the covariate followed a Bernoulli distribution.

Among various sample size methods for logistic regression, the main target is to find an appropriate variance estimator for the test statistics. The literature has demonstrated that the maximum likelihood method for estimating the variance of the estimated regression coefficients is not simple. The two group comparison methods by Lachin (1981) and Donner (1984) did not account for the influence of multiple covariates. In the next chapter, we propose an analogous sample size study from Vittinghoff et al. (2012, p.74, p.130, p.194) for the RR. The idea of using the Wald test by Demidenko (2007) is followed in the sample size formulation. We also implement the VIF in the next chapter.

Chapter 3 Sample Size for Modified Poisson Regression

This chapter applies the Wald test statistics, the least square estimation, and the Delta method to derive a sample size formula for risk ratios. As Vittinghoff et al. (2012, p.74, p.130, p.194) derived the OR sample size based on the least-square method, we also start the RR sample size derivation with the same method. Following the general principle from Chapter 2, we implement the variance inflation factor to account for multicollinearity among multiple covariates. We assume that the parameter estimate of interest is $\widehat{\beta_1}$ in the multiple regression, representing the log of RR or regression coefficient of the risk factor x_1.

3.1. Derivation of variance

If there is no difference in the probability of the outcome event between groups $x_1 = 1$ and $x_1 = 0$, the Wald test statistic for the null hypothesis of no effect is given by

$$\text{Wald} = \frac{\left(\widehat{\beta_1} - 0\right)^2}{\text{Var}\left(\widehat{\beta_1}\right)}. \tag{3.1}$$

To find the appropriate estimator and its variance, we first apply the least-square method on the linear regression. Let **X** and **Y** represent the covariate matrix and outcome vector, $\boldsymbol{\varepsilon}$ be a vector of the true unobserved residuals, and $\boldsymbol{\beta}$ be a vector of unknown population parameters. The simple version of the matrix equation is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. The estimated minimum sum of square of residuals is $\boldsymbol{e'e} = \left(\boldsymbol{Y} - \boldsymbol{X\widehat{\beta}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X\widehat{\beta}}\right)$, where $\boldsymbol{\widehat{\beta}}$ is the estimate of $\boldsymbol{\beta}$, $\boldsymbol{e}$ is a vector of residuals that can be observed, and $\boldsymbol{e'e}$ is a scalar or a number. Taking the derivative of $\boldsymbol{e'e}$ with respect to $\boldsymbol{\widehat{\beta}}$, and $\boldsymbol{\widehat{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ after setting the derivative equation to 0.

Assuming homoskedasticity (constant residual variance) and no correlation in the unobserved , we obtain $E(\boldsymbol{\varepsilon'\varepsilon}|\boldsymbol{X}) = \sigma_\varepsilon^2\boldsymbol{I}$, where σ_ε^2 is the residual variance and $\boldsymbol{I}$ is the identity matrix. The expected variance covariance matrix of $\boldsymbol{\widehat{\beta}}$ can be calculated as

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \mathrm{E}\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_\varepsilon^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \qquad (3.2)\end{aligned}$$

The variance of the unbiased estimator $\widehat{\beta_1}$ is

$$\begin{aligned}\mathrm{Var}(\widehat{\beta_1}) &= \frac{n\sigma_\varepsilon^2}{n\sum x_{1i}^2 - (\sum x_{1i})^2} \\ &= \frac{\sigma_\varepsilon^2}{\sum(x_{1i} - \bar{x}_1)^2} \\ &= \frac{\sigma_\varepsilon^2}{(n-1)\sigma_{x_1}^2},\end{aligned}$$

where x_{1i} is the i^{th} element of the x_1 vector ($i \in (1,2,\dots n)$), $\bar{x}_1$ is the average of the x_1 variable, and $\sigma_{x_1}^2$ represents the variance of the x_1.

When there is only one covariate, the relationship among the variance of the response variable σ_y^2, the variance of residuals σ_ε^2, and the coefficient of determination $r_{y,1}^2$ can be expressed as (Hsieh et al., 2003)

$$\sigma_\varepsilon^2 = \sigma_y^2(1 - r_{y,1}^2), \qquad (3.4)$$

where $r_{y,1}^2$ shows the proportion to which the covariate x_1 can explain the variation of response in the regression model. If there is no covariate in the model, the $r_{y,1}^2$ will become zero, and the variation of the response variable is equal to the variance of residuals. If there is more than one covariate in the regression model, the VIF is applied for adjustment of multicollinearity among covariates in replacement of $(1 - r_{y,1}^2)$. The VIF is expressed as $1/(1 - r_{1,2,\dots,k}^2)$, where $r_{1,2\dots k}^2$ is the coefficient of determination from regressing x_1 on other covariates $x_2, \dots, x_k$. We can extend the equation as

$$\sigma_\varepsilon^2 = \frac{\sigma_y^2}{1 - r_{1,2\dots k}^2}. \qquad (3.5)$$

In the modified Poisson regression model, the probability of the outcome event is modeled as a function of covariates using a log link. Let p be the overall prevalence for the binary outcome, the variance of the log(p) can be obtained using Delta method

$$\begin{aligned}\sigma_y^2 &= \mathrm{Var}\left(\log(p)\right)\\ &= \left(\frac{\partial\log(p)}{\partial p}\right)^2 \mathrm{Var}(p)\\ &= \left(\frac{1}{p}\right)^2 p(1-p)\\ &= \frac{1-p}{p}.\end{aligned} \tag{3.6}$$

By replacing the σ_ε^2 and σ_y^2 and using n to approximate $n-1$, the variance of $\widehat{\beta_1}$ can be obtained by accommodating the influence of having multiple covariates:

$$\mathrm{Var}\left(\widehat{\beta_1}\right) = \frac{1}{n\sigma_{x_1}^2}\frac{1-p}{p}\frac{1}{1-r_{1,2\ldots k}^2}. \tag{3.7}$$

3.2. Sample size, power, and minimal detectable effect

Now we derive the formulas for sample size, power and minimal detectable effect for testing risk ratios in the modified Poisson regression model. The minimum detectable effect is the smallest value of the regression coefficient of interest, provided the sample size and power to reject the null hypothesis (Vittinghoff et al., 2012, p.131). We refer to the general principle for finding a sample size by using the power calculation of a hypothesis test. The null hypothesis is when the parameter of interest $\beta_1 = 0$, and the alternative hypothesis is $\beta_1 \neq 0$. The Wald test statistic is used in test. Assuming the parameter of interest under the alternative hypothesis is positive, $\Pr\left(\widehat{\beta_1}/\mathrm{SE}\left(\widehat{\beta_1}\right) < Z_{\alpha/2} | \mathrm{H}_1\right) \approx 0$. Let β_1^* be the value of the population parameter of interest under the alternative hypothesis, we add an extra term of $\beta_1^*/\mathrm{SE}\left(\widehat{\beta_1}\right)$ to the both sides of the inequality in Equation (3.8) and obtain the power of the Wald test as

$$\gamma \approx \Pr\left(\frac{\widehat{\beta_1} - 0}{\mathrm{SE}\left(\widehat{\beta_1}\right)} > Z_{1-\alpha/2}\middle| \mathrm{H}_1\right) \tag{3.8}$$

$$= \Pr\left(\frac{\widehat{\beta_1}}{\mathrm{SE}\left(\widehat{\beta_1}\right)} - \frac{\beta_1^*}{\mathrm{SE}\left(\widehat{\beta_1}\right)} > Z_{1-\alpha/2} - \frac{\beta_1^*}{\mathrm{SE}\left(\widehat{\beta_1}\right)}\right),$$

where $\left(\widehat{\beta_1} - \beta_1^*\right)/\mathrm{SE}\left(\widehat{\beta_1}\right)$ follows a standard normal Z-distribution. Let Φ be the cumulative distribution function of the standard normal distribution. Replacing the $\mathrm{SE}\left(\widehat{\beta_1}\right)$ with the square root of the estimated variance, the sample size is given by

$$\gamma = \Phi\left(\frac{\beta_1^*}{\mathrm{SE}\left(\widehat{\beta_1}\right)} - Z_{1-\alpha/2}\right)$$

$$\frac{\beta_1^*}{Z_{1-\alpha/2} + Z_\gamma} = \mathrm{SE}\left(\widehat{\beta_1}\right)$$

$$\frac{(\beta_1^*)^2}{\left(Z_{1-\alpha/2} + Z_\gamma\right)^2} = \frac{1}{n\sigma_{x_1}^2}\frac{1-p}{p}\frac{1}{1-r_{1,2\ldots k}^2}$$

$$n = \frac{\left(Z_{1-\alpha/2} + Z_\gamma\right)^2}{(\beta_1^*)^2\sigma_{x_1}^2}\frac{1-p}{p}\frac{1}{1-r_{1,2\ldots k}^2}. \tag{3.9}$$

The power and minimal detectable effect are obtained as

$$\gamma = 1 - \Phi\left(Z_{1-\alpha/2} - |\beta_1^*|\sigma_{x_1}\sqrt{n\frac{p}{1-p}\left(1 - r_{1,2\ldots k}^2\right)}\right), \tag{3.10}$$

$$\pm\beta_1^* = \frac{Z_{1-\alpha/2} + Z_\gamma}{\sigma_{x_1}\sqrt{n\frac{p}{1-p}\left(1 - r_{1,2\ldots k}^2\right)}}. \tag{3.11}$$

Equations (3.9) to (3.11) are derived formulas for sample size, power and minimum detectable effect that account for the influence of covariates other than the factor of interest. All derivations are based on asymptotics. We conduct a simulation study to evaluate the performance of the power formula in Equation (3.10) in Chapter 4.

Chapter 4 Simulation Study

In this chapter, we conduct a simulation study to evaluate the proposed sample size formula for the modified Poisson model under practical situations. The main objectives of the simulation study are:

1. To evaluate the empirical power of the modified Poisson model for detecting the effect of a risk factor when the response is generated from the logistic model.
2. To examine the performance of the proposed sample size and power formulas for the modified Poisson model.

4.1. Simulation design and data generation

The simulation study consists of two parts. The first part compares the empirical powers from the logistic regression and modified Poisson model when the response probabilities are generated from the logistic model. The second part assesses the adequacy of the proposed sample size formula for the modified Poisson model.

In each part, we consider five scenarios: two scenarios are with a single covariate, and the other three scenarios with two covariates. With a single covariate x, the covariate is considered to be binary or continuous. With two-covariates (x_1, x_2), the scenarios are based on two binary covariates, two continuous covariates, or a mixture of one binary (x_1) and one continuous (x_2) covariate, where the risk factor of interest is x_1. The two-covariate scenarios are designed to evaluate the power of detecting the effect of the risk factor when adjusting for the other covariate x_2. In the two-covariate scenarios, we denote by r the correlation between the two covariates. Then the coefficient of determination obtained from the model of x_1 regressing on x_2 is expressed as r^2.

As SAS® is widely used in the pharmaceutical industry, it is used as the tool for the simulation study. In Sections 4.1.1 to 4.1.3, we describe the detailed steps of the data-generating and simulation processes.

4.1.1. Generating covariates

Prior to generating response variables, we generate the covariates. A single covariate, x is generated using the RAND function from SAS 9.4 with Bernoulli or normal distribution. For the two-covariate scenarios, the correlated covariates are generated based on the RandMVBinary, RandNormal, and Cholesky transformation from SAS/IML software (Wicklin, 2013, pp. 133-157, 176-177). The RandMVBinary function generates multivariate binary data. The RandNormal function in SAS/IML is used for simulating correlated multivariate normal data with a predetermined mean and covariance matrix. In scenarios with one binary and one continuous covariate, the data is simulated using the Cholesky transformation. The Cholesky transformation uses a given covariance structure to simulate multivariate normal data. The simulated data for x_1 is converted to binary, and $x_1 = 1$ if the generated value is greater than zero. There is no additional adjustment applied to the correlation reduction due to the dichotomization in the generated data.

The binary covariates were generated with the probability of 0.5 (exposed = 1 and unexposed = 0). The continuous covariates were generated from a standard normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$. The correlation coefficients used for generating two covariates were 0, 0.3, and 0.6 to represent no, medium and large correlations (Cohen, 1988, pp. 79-81).

4.1.2. Simulation settings for empirical power comparison

In the simulation study for the first objective to compare empirical power of the modified Poisson model to that of the logistic model, we considered the total sample size of 300 and first generated the covariates with pre-specified probability of 0.5 for binary covariates or mean of 0 and variance of 1 for continuous covariates and correlation coefficient r of $0, 0.3, 0.6$ for two covariate scenarios. Then, we generated the response variables based on the response probability obtained from the logstic regression model with fixed covariates, odds ratio values of $OR_{x_1} = 1, 1.5, 2,$ and 2.5 for the risk factor of interest x_1, $OR_{x_2} = 1, 1.5,$ and 2 for x_2, and the baseline response probability p_0 of $0.1, 0.2, 0.3,$ and 0.4 for one covariate scenarios or 0.1 and 0.4 for two covariate scenarios, where $p_0 = \Pr(Y|x = 0)$ or $p_0 = \Pr(Y|x_1 = x_2 = 0)$. We set the significance level α to be 0.05.

To be more specific, pre-specified odds ratios, baseline response probability and fixed values of covariates were considered, the individual response probability was calculated from the logistic regression model as: $p_i = e^{\beta_0+\beta_1 x_i}/\left[1+e^{\beta_0+\beta_1 x_i}\right]$ for one-covariate scenarios or $p_i = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}/\left[1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}\right]$ for two-covariate scenarios. The β_0 was obtained as the log of $p_0/(1-p_0)$ and the β_1 as the log of OR_x or OR_{x_1}, and β_2 as the log of OR_{x_2}. The individual response y_i was simulated with a random Bernoulli function given p_i.

4.1.3. Simulation settings for sample size examination

The simulation setting for evaluating the proposed risk ratio power formula for the modified Poisson model is similar to that was described in the previous section. The pre-specified parameters, the generated covariates, and the α value remained the same. The sample sizes were fixed as 300 and 500.

For second part of the simulation, the modified Poisson model was used in generating the responses. The response probabilities p_i were obtained by using anti-log link, preset RRs, and given covariate values. Individual p_i equaled $e^{\beta_0+\beta_1 x_i}$ or $e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}$. The β_0 was the log of p_0, where $p_0 = \Pr(Y|x=0)$ or $p_0 = \Pr(Y|x_1 = x_2 = 0)$. The baseline response probability p_0 varied at 0.1, 0.2, 0.3, and 0.4. The RR_x or RR_{x_1} for x or x_1 varied at 1, 1.5, 2, and 2.5, $RR_{x_2} \in (1, 1.5, 2)$, and the regression coefficients β_1 and β_2 were determined by the logarithms of RRs.

Because an anti-log link was used to obtain the response probability, it could be greater than one. For example, in scenarios with two binary covariates, when $p_0 = 0.3, RR_{x_1} = RR_{x_2} = 2.5$ and $x_1 = x_2 = 1$, the corresponding response probability becomes greater than one, thus, we excluded this parameter value combination in the simulation. For the scenarios of one binary or two binary covariates, we considered only workable parameter value combinations that make the response probability less than one. For the scenarios involving a continuous covariate, following Yelland et al. (2011), the covariate values were regenerated when the individual p_i was greater than one to avoid the response generation errors.

4.2. Assessment criteria

For each simulation setting, we evaluated the power of the modified Poisson model using 1000 simulations of a given sample size in terms of empirical powers comparing to nominal powers. The same number of runs was used in similar simulation studies (Hsieh et al., 1998; Zou, 2004). Each run had the same generated covariate values, but individual y_i varied. The Genmod procedure from SAS 9.4 was used to fit the logistic regression and modified Poisson models to analyze the same data set in each run. For both models, the null hypothesis of testing the significance of the regression coefficient is $\mathrm{H}_0 : \beta_1 = 0$. The powers of the two models are comparable since the null hypothesis is identical. The empirical power in both parts of the simulation study was calculated as the times of rejecting the null hypothesis out of 1000 simulation runs when the P-value < 0.05. In part one, if the empirical powers from the logistic regression and the modified Poisson model have a less than 5% difference, we consider the two powers are equivalent.

In the first part of simulation, the nominal powers for detecting the odds ratios of a given sample size were obtained for the logistic regression model for comparison using the following formula (Vittinghoff et al., 2012, p. 195):

$$\gamma = 1 - \Phi\left(Z_{1-\alpha/2} - |\beta_1^*| \sigma_{x_1} \sqrt{np(1-p)\left(1 - r_{1,2\ldots k}^2\right)} \right). \qquad (4.1)$$

We used the weighted average as the overall prevalence p for the outcome when the covariate was binary, with weight as the group size of $x = 1$ or $x = 0$, since using the weighted average is more accurate than using an arithmetic average. Weight can be calculated as the number of $x = 1$ or $x = 0$ divided by *n*. For a balanced design, equal weights were used. For one binary covariate scenario, the weight was 0.5. For the two binary covariate scenario, since $\Pr(x_1 = 1) = \Pr(x_2 = 1) = 0.5$, the weight for each x_1 and x_2 combined category was 0.25. In the scenarios with a continuous covariate, the average of 300 individual p_i was used as p. The standard deviation of the binary covariate of interest was calculated as the square root of $\Pr(x_1 = 1)\left(1 - \Pr(x_1 = 1)\right)$. For the interested continuous covariate, the standard deviation was $\sqrt{\sigma^2} = 1$. The other values required to find the nominal power were pre-specified. For one-covariate scenarios, the x_1 in the Equation (4.1) represents the only covariate x, and the squared correlation coefficient $r^2_{1,2\dots k}$ was not included in calculating the nominal power. For other scenarios, the correlation coefficient in the power equation was a preset value of 0, 0.3 or 0.6. In practice, the correlation coefficient for two covariates can be estimated by the Pearson correlation in the sample data.

In the second part, the proposed power formula was used to calculate the nominal power for detecting risk ratios. The weighted average or the mean of p_i was used as the overall prevalence p in different scenarios. The standard deviation of the binary covariate was calculated as 0.5, since the probability for x or $x_1 = 1$ was set as 0.5. The standard deviation of the continuous covariate equaled $\sigma = 1$. The other parameter values applied in the proposed power formula were preset from Section 4.1.3.

For comparing the empirical power of the logistic regression to its nominal power and the empirical power of the modified Poisson model to the nominal power for detecting RR, we consider the power difference is acceptable if the difference between the nominal and empirical powers is less than 10%.

4.3. Simulation results

We summarize the simulation results for the two objectives in separate sections. Section 4.3.1 evaluates the empirical powers for the modified Poisson model compared to those from the logistic regression. Section 4.3.2 assesses the proposed sample size and power formulas for the modified Poisson model comparing their nominal and empirical powers.

4.3.1. Empirical power examination

The simulation results to evaluate the first objective of the simulation study are summarized in Table 4.1 to 4.5. The power of detecting the effect of a risk factor using the modified Poisson was evaluated when the response data were generated from the logistic model under different scenarios. For all simulation settings, the significance level was set as $\alpha = 0.05$ and a two-sided test was used.

4.3.1.1. Power for a binary risk factor

The simulation results for one binary covariate scenario are summarized in Table 4.1 to compare the power of the modified Poisson model to that of the logistic regression. In general, we observed that the differences between the empirical powers of the two models are negligible regardless the effect size and the baseline response probability p_0,. The power of the logistic model was closely estimated to the nominal power. As expected, the nominal and empirical powers increase as OR_x increases when p_0 is fixed, and also as the p_0 increases when OR_x is fixed.

Table 4.2 presents the simulation results for the power of detecting the effect of the risk factor of interest when adjusting for a binary covariate x_2. The same results are visually illustrated in Figure 4.1 for each OR_{x_2}, p_0 and r combinations when OR_{x_1} varies. As expected, the simulation results with $OR_{x_2} = 1, r = 0$ are similar to those from the one-covariate scenario in Table 4.1. The empirical powers from the logistic regression and the modified Poisson model agree with each other for all combinations of $p_0, OR_{x_1}, OR_{x_2}, r$. The differences between the two empirical powers are negligible, although the empirical power from the modified Poisson model is slightly lower than that from the logistic regression when the correlation between the two covariates is high. The difference between the nominal and empirical powers of the logistic regression is small. When r reaches 0.6, the empirical power is slightly higher than the nominal power.

Likewise, when adjusting for a continuous covariate x_2, Table 4.3 represents the simulation results under the same parameter setups used for a binary covariate adjustment as shown in Table 4.2. Figure 4.2 graphically displays the power trend using the results from Table 4.3. The differences between empirical powers from the logistic regression and modified Poisson model are negligible. The empirical power from the logistic regression is close to its nominal power in most of settings except when the correlation is large. The difference between the nominal and empirical powers was ranged between (0%, 7.4%), and the highest difference occurred at $OR_{x_1} = 2.5, OR_{x_2} = 2, r = 0.6, p_0 = 0.1$. When the correlation increases, the empirical power was decreased due to a multicollinearity problem. The power trend of varying OR_{x_1} for the scenario of one binary and one continuous covariates shows the minor difference between the nominal power for estimating OR and the empirical power of the logistic regression. The empirical power of the logistic regression is slightly lower than the nominal power when OR_{x_2} and r increase. The empirical powers of the logistic regression and modified Poisson model remain close to each other for every OR_{x_2} and r combinations.

Table 4.1 Results of the empirical powers from the logistic regression ($EP_{logistic}$) and the modified Poisson model ($EP_{mpoisson}$) for one binary covariate. The odds ratio $OR_x \in (1,1.5,2,2.5)$, the response probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. The nominal powers (NP) are obtained from the logistic model. Empirical powers are estimated based on 1000 runs, $n = 300$.

p_0	OR_X	p	NP	$EP_{logistic}$	$EP_{mpoisson}$
0.1	1	0.10	5.0	4.1	4.1
	1.5	0.12	20.8	21.3	20.9
	2	0.14	55.1	53.4	52.5
	2.5	0.16	82.6	79.8	79.7
0.2	1	0.20	5.0	4.7	4.3
	1.5	0.24	31.9	30.9	30.8
	2	0.27	75.6	73.8	73.8
	2.5	0.29	95.0	94.2	94.2
0.3	1	0.30	5.0	3.7	3.6
	1.5	0.35	38.5	36.4	36.4
	2	0.38	83.0	82.8	82.5
	2.5	0.41	97.4	97.0	97.0
0.4	1	0.40	5.0	4.1	4.0
	1.5	0.45	41.5	41.7	41.3
	2	0.49	85.1	83.8	83.7
	2.5	0.51	97.8	98.1	98.1

Table 4.2 Results of the empirical powers from the logistic regression ($EP_{logistic}$) and the modified Poisson model ($EP_{mpoisson}$) for two binary covariates. The odds ratios $OR_{x_1} \in (1,1.5,2,2.5), OR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the response probability of the baseline group $p_0 \in (0.1,0.4)$, and p represents the overall prevalence. The nominal powers (NP) are obtained from the logistic model. Empirical powers are estimated based on 1000 runs, $n = 300$.

			$p_0 = 0.1$				$p_0 = 0.4$			
$ORx1$	$ORx2$	r	p	NP	$EP_{logistic}$	$EP_{mpoisson}$	p	NP	$EP_{logistic}$	$EP_{mpoisson}$
1	1	0	0.10	5.0	4.8	4.7	0.40	5.0	4.5	4.3
1.5	1	0	0.12	20.8	19.2	18.5	0.45	41.5	41.6	41.0
		0.3	0.12	19.3	19.0	18.6	0.45	38.4	39.4	39.0
		0.6	0.12	14.8	15.5	16.5	0.45	28.6	30.8	30.3
1.5	1.5	0	0.15	23.6	24.4	23.3	0.50	41.8	41.8	41.4
		0.3	0.15	21.9	21.3	20.7	0.50	38.7	39.4	38.3
		0.6	0.15	16.6	16.3	17.0	0.50	28.9	31.1	29.5
1.5	2	0	0.17	25.9	27.5	25.7	0.54	41.7	41.1	41.1
		0.3	0.17	24.0	24.2	23.2	0.54	38.6	40.1	38.5
		0.6	0.17	18.2	18.5	17.6	0.54	28.8	31.5	29.7
2	1	0	0.14	55.1	56.0	54.5	0.49	85.1	85.0	84.2
		0.3	0.14	51.3	51.8	51.8	0.49	81.6	82.9	82.5
		0.6	0.14	38.6	40.2	41.7	0.49	67.0	71.4	71.2
2	1.5	0	0.17	61.3	61.9	61.5	0.54	84.9	84.2	84.1
		0.3	0.17	57.3	57.6	56.8	0.54	81.5	82.3	81.5
		0.6	0.17	43.6	45.7	45.0	0.54	66.8	71.6	69.9
2	2	0	0.19	65.8	67.0	66.2	0.57	84.5	84.3	84.3
		0.3	0.19	61.7	62.3	60.3	0.57	81.0	80.6	79.1
		0.6	0.19	47.4	50.6	47.5	0.57	66.2	70.8	68.0
2.5	1	0	0.16	82.6	82.4	81.8	0.51	97.8	97.9	97.7
		0.3	0.16	79.0	79.9	78.7	0.51	96.6	97.0	96.9
		0.6	0.16	64.0	68.3	68.3	0.51	88.7	91.5	91.3
2.5	1.5	0	0.19	87.3	86.4	85.9	0.56	97.6	96.9	96.9
		0.3	0.19	84.1	84.3	83.8	0.56	96.4	96.4	96.1
		0.6	0.19	69.9	73.4	71.4	0.56	88.3	91.4	90.2
2.5	2	0	0.21	90.2	90.7	90.5	0.59	97.4	97.0	97.0
		0.3	0.21	87.4	87.5	86.4	0.59	96.1	96.7	96.2
		0.6	0.21	74.0	78.0	75.7	0.59	87.7	91.3	90.3

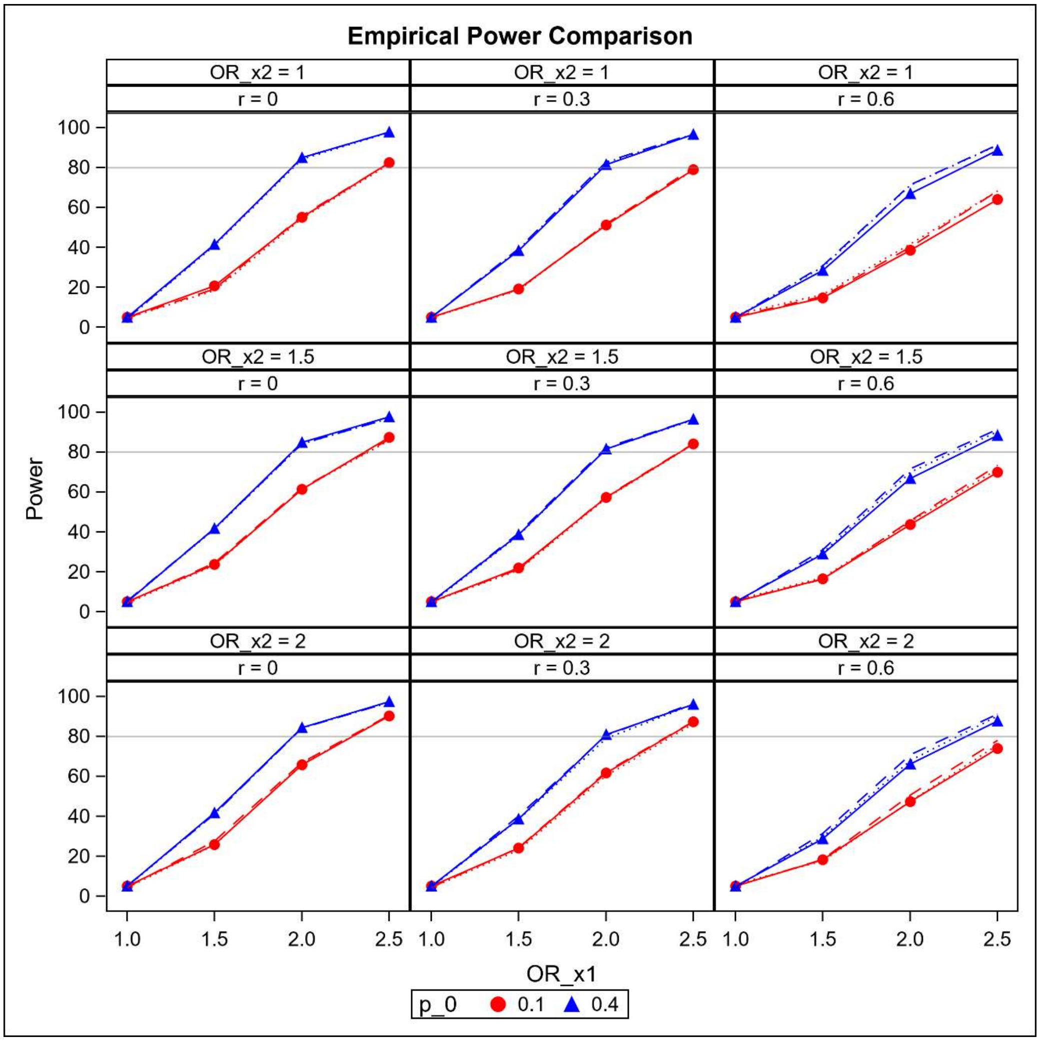

Figure 4.1 Comparison of nominal and empirical powers of testing for the effect of a binary risk factor x_1 adjusting for a binary covariate x_2 based on simulation results from 1000 simulations of a sample size of 300 as shown in Table 4.2. The OR_{x_1} and OR_{x_2} represent the odds ratios for x_1 and x_2, respectively, r represents the correlation between the two covariates, p_0 is the probability of outcome when both covariates are zero. Solid line represents the nominal power of the logistic regression; dash line represents the empirical power of the logistic regression; dotted line represents the empirical power of the modified Poisson model.

Table 4.3 Results of the empirical powers from the logistic regression ($EP_{logistic}$) and the modified Poisson model ($EP_{mpoisson}$) for one binary and one continuous covariates. The odds ratios $OR_{x_1} \in (1,1.5,2,2.5), OR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the response probability of the baseline group $p_0 \in (0.1,0.4)$, and p represents the overall prevalence. The nominal powers (NP) are obtained from the logistic model. Empirical powers are estimated based on 1000 runs, $n = 300$

			p_0 = 0.1				p_0 = 0.4			
ORx1	*ORx2*	*r*	*p*	*NP*	*EPlogistic*	*EPmpoisson*	*p*	*NP*	*EPlogistic*	*EPmpoisson*
1	1	0	0.10	5.0	4.1	4.1	0.40	5.0	4.1	4.0
1.5	1	0	0.12	20.7	21.4	21.2	0.45	41.5	41.8	41.5
		0.3	0.12	19.2	19.6	19.5	0.45	38.4	38.3	37.8
		0.6	0.12	14.8	15.2	16.2	0.45	28.6	27.7	28.5
1.5	1.5	0	0.13	21.4	21.4	21.3	0.45	41.5	38.2	38.1
		0.3	0.13	20.1	18.9	17.8	0.45	38.4	36.3	35.2
		0.6	0.13	15.5	14.8	16.1	0.45	28.6	28.0	27.5
1.5	2	0	0.14	22.7	21.0	20.8	0.45	41.5	38.2	38.5
		0.3	0.14	21.4	19.4	17.8	0.45	38.4	34.1	32.8
		0.6	0.15	16.5	14.9	15.9	0.45	28.7	25.3	26.8
2	1	0	0.14	54.6	53.4	52.7	0.48	85.1	83.7	83.7
		0.3	0.14	50.8	49.6	48.8	0.48	81.6	79.5	79.5
		0.6	0.14	38.2	37.2	37.9	0.48	67.0	66.3	66.0
2	1.5	0	0.15	56.2	52.7	53.1	0.48	85.1	82.1	82.2
		0.3	0.15	53.1	50.4	48.8	0.48	81.6	79.4	78.0
		0.6	0.15	40.8	38.0	37.9	0.48	67.0	63.9	64.3
2	2	0	0.16	58.9	55.4	55.0	0.48	85.1	81.4	81.9
		0.3	0.16	56.1	52.0	49.7	0.48	81.6	77.0	75.5
		0.6	0.17	43.6	39.3	40.8	0.48	67.0	64.8	66.4
2.5	1	0	0.16	82.1	79.8	79.8	0.51	97.8	98.1	98.1
		0.3	0.16	78.4	77.6	77.9	0.51	96.6	96.9	96.7
		0.6	0.16	63.4	60.7	61.6	0.51	88.7	88.0	87.9
2.5	1.5	0	0.16	83.3	80.4	80.0	0.51	97.8	96.7	96.7
		0.3	0.17	80.6	78.2	77.1	0.51	96.6	96.0	95.3
		0.6	0.17	66.8	62.9	62.5	0.51	88.7	87.2	87.0
2.5	2	0	0.17	85.3	82.2	82.2	0.51	97.8	97.1	97.2
		0.3	0.18	83.1	79.3	78.0	0.51	96.6	96.1	95.4
		0.6	0.19	70.1	62.7	65.4	0.51	88.7	85.6	87.5

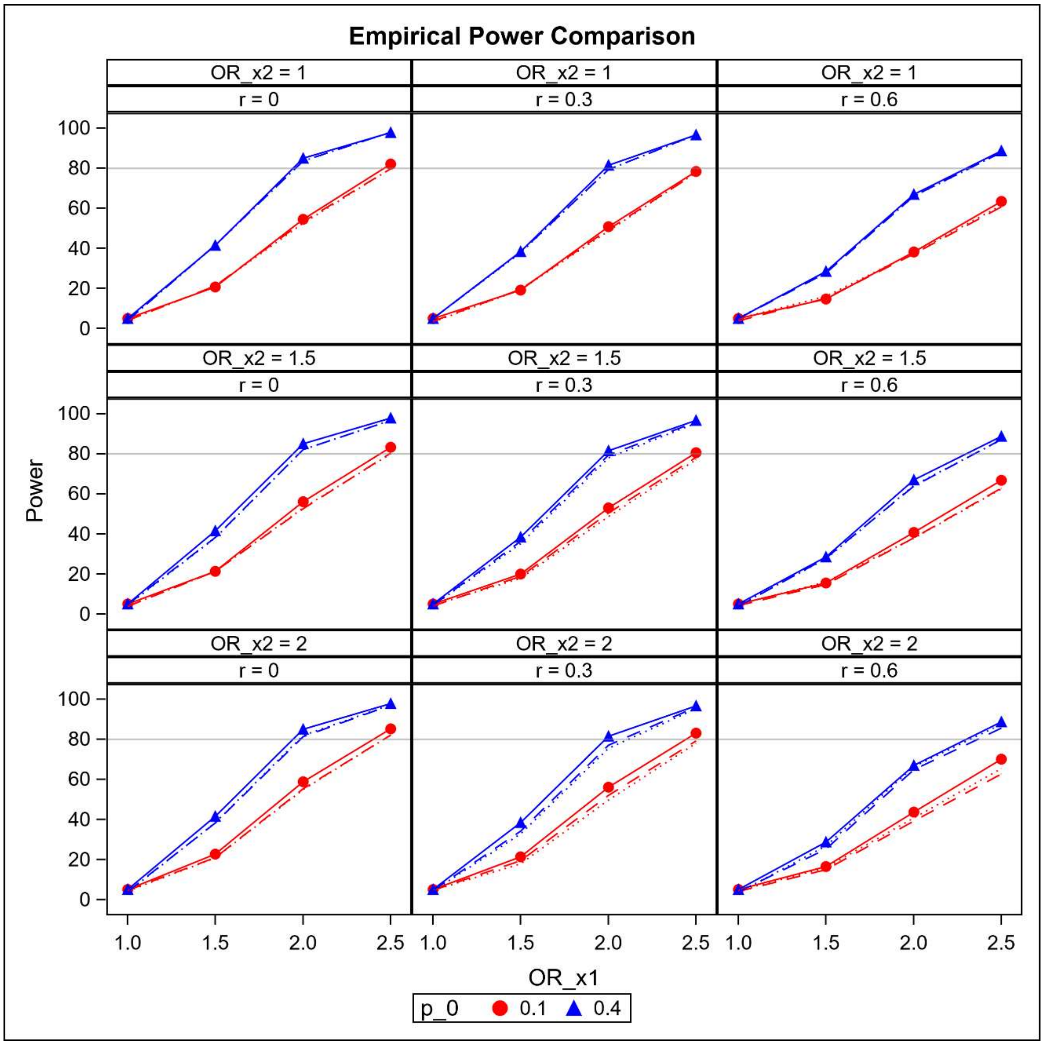

Figure 4.2 Comparison of nominal and empirical powers of testing for the effect of a binary risk factor x_1, adjusting for a continuous covariate x_2 based on the simulation results from 1000 simulations of a sample size of 300 as shown in Table 4.3. The OR_{x_1} and OR_{x_2} represent the odds ratios for x_1 and x_2, respectively, r represents the correlation between the two covariates, p_0 is the probability of outcome when both covariates are zero. Solid line represents the nominal power of the logistic regression; dash line represents the empirical power of the logistic regression; dotted line represents the empirical power of the modified Poisson model.

4.3.1.2. Power for a continuous risk factor

Table 4.4 presents the simulation results of the nominal power for detecting OR and the empirical powers of the logistic regression and the modified Poisson model under one continuous covariate scenario. The greatest difference between the two empirical powers is 2.3%. The nominal power and the two empirical powers rise more quickly within the OR_x range compared to the powers from one binary risk factor scenario. One of the reasons for this phenomenon could be that the variance for the continuous variable is larger than that of the binary variable. Another reason could be that the continuous variable contains more information than a binary variable, it is reasonably to have a higher power when the risk factor is a continuous variable (Altman & Royston, 2006). The powers are greater than 95% when $OR_x \geq 2$ for the listed p_0.

Having a continuous factor of interest and adjusting for a continuous covariate x_2, all setups have the two empirical powers close to each other in Table 4.5. The largest difference is 5.6%, which is marginally higher than our power equivalency standard. As comparison, the nominal powers are compared to the empirical powers based on logistic regression models. The nominal and empirical powers are similar to each other in most settings except for the settings with $OR_{x_1} = 1.5, OR_{x_2} > 1, r = 0.6, p_0 = 0.4$, in which the nominal power was overestimated greater than 10% than the empirical power.

We note that the settings of different combinations of p_0, OR_{x_1}, OR_{x_2}, and r did not yield any warnings during simulations in Tables 4.1 to 4.5. There were some settings with 100.0% power for the two continuous covariates scenario. The higher power could be caused by the large effect sizes used in the simulation. Another reason could be that the sample size was larger than the needed n. For example, in Table 4.5, with $p_0 = 0.1, \text{O}R_{x_1} = \text{O}R_{x_2} = 2, r = 0$, reduced sample size of $n = 200$ provided 92.1% nominal power, and 86.8% and 90.8% empirical powers for the logistic regression and the modified Poisson models, respectively. In general, the empirical powers from both regression models were close to each other in various scenarios.

Table 4.4 Results of the empirical powers from the logistic regression ($EP_{logistic}$) and the modified Poisson model ($EP_{mpoisson}$) for one continuous covariate. The odds ratio $OR_x \in (1,1.5,2,2.5)$, the response probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. The nominal powers (NP) are obtained from the logistic model. Empirical powers are estimated based on 1000 runs, $n = 300$.

p_0	OR_X	p	NP	$EP_{logistic}$	$EP_{mpoisson}$
0.1	1	0.10	5.0	4.3	5.5
	1.5	0.11	58.6	58.0	60.3
	2	0.12	97.4	98.2	97.8
	2.5	0.13	100.0	100.0	100.0
0.2	1	0.20	5.0	3.9	5.2
	1.5	0.21	81.7	79.0	81.1
	2	0.23	99.9	99.8	99.8
	2.5	0.24	100.0	100.0	100.0
0.3	1	0.30	5.0	4.5	5.0
	1.5	0.31	90.1	88.0	88.8
	2	0.32	100.0	100.0	100.0
	2.5	0.34	100.0	100.0	100.0
0.4	1	0.40	5.0	4.6	5.2
	1.5	0.41	93.2	90.6	91.4
	2	0.42	100.0	100.0	100.0
	2.5	0.42	100.0	100.0	100.0

Table 4.5 Results of the empirical powers from the logistic regression ($EP_{logistic}$) and the modified Poisson model ($EP_{mpoisson}$) for two continuous covariates. The odds ratios $OR_{x_1} \in (1,1.5,2,2.5), OR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the response probability of the baseline group $p_0 \in (0.1,0.4)$, and p represents the overall prevalence. The nominal powers (NP) are obtained from the logistic model. Empirical powers are estimated based on 1000 runs, $n = 300$

			p_0 = 0.1				p_0 = 0.4			
ORx1	*ORx2*	*r*	*p*	*NP*	*EPlogistic*	*EPmpoisson*	*p*	*NP*	*EPlogistic*	*EPmpoisson*
1	1	0	0.10	5.0	3.7	5.8	0.40	5.0	3.9	4.9
1.5	1	0	0.11	57.9	52.4	56.9	0.40	93.1	90.8	91.1
		0.3	0.11	53.9	46.7	50.5	0.40	90.7	85.7	86.6
		0.6	0.11	40.8	33.8	36.0	0.40	78.6	70.6	72.1
1.5	1.5	0	0.11	59.6	54.1	**59.4	0.41	93.1	89.9	91.1
		0.3	0.11	56.8	51.5	54.2	0.41	90.8	84.0	86.4
		0.6	0.12	44.0	35.8	36.1	0.41	*78.8	67.6	68.1
1.5	2	0	0.12	62.8	54.8	**60.4	0.41	93.2	86.8	88.5
		0.3	0.13	60.4	50.6	54.7	0.41	90.9	82.8	84.2
		0.6	0.13	47.5	37.8	36.6	0.42	*79.0	64.7	61.5
2	1	0	0.12	97.1	95.0	95.6	0.41	100.0	100.0	100.0
		0.3	0.12	95.7	91.7	93.3	0.41	100.0	99.9	100.0
		0.6	0.12	86.9	80.2	82.7	0.41	99.7	99.0	99.2
2	1.5	0	0.12	97.5	94.7	96.1	0.41	100.0	100.0	100.0
		0.3	0.13	96.8	92.7	94.1	0.42	100.0	99.8	99.8
		0.6	0.13	90.1	81.8	81.7	0.42	99.7	98.3	98.0
2	2	0	0.13	98.2	95.9	96.6	0.42	100.0	100.0	100.0
		0.3	0.14	97.7	93.6	94.8	0.42	100.0	99.6	99.9
		0.6	0.15	92.5	82.6	79.1	0.42	99.7	97.3	96.2
2.5	1	0	0.13	100.0	99.6	99.7	0.42	100.0	100.0	100.0
		0.3	0.13	100.0	99.5	99.5	0.42	100.0	100.0	100.0
		0.6	0.13	98.8	97.0	96.8	0.42	100.0	100.0	100.0
2.5	1.5	0	0.13	100.0	99.8	99.8	0.42	100.0	100.0	100.0
		0.3	0.14	99.9	99.8	99.9	0.42	100.0	100.0	100.0
		0.6	0.14	99.4	96.9	95.8	0.42	100.0	100.0	99.9
2.5	2	0	0.14	100.0	99.7	99.7	0.42	100.0	100.0	100.0
		0.3	0.15	100.0	99.9	99.9	0.43	100.0	100.0	100.0
		0.6	0.16	99.6	97.5	95.6	0.43	100.0	100.0	99.7

* The difference between NP and $EP_{logistic} >= 10\%$

** The difference between $EP_{logistic}$ and $EP_{mpoisson} \geq 5\%$

4.3.2. Performance of sample size formulas for estimating risk ratios

In this section, we summarize the simulation results to evaluate the proposed power formula for the modified Poisson model in Tables 4.6 to 4.10. The nominal and empirical powers of detecting the effect of a risk factor based on the modified Poisson were evaluated when the response data were generated from the modified Poisson model. The sample sizes were set at 300 and 500.

4.3.2.1. Power for a binary risk factor

We first evaluated the proposed RR power formula for the modified Poisson model with one binary covariate. The simulated results are summarized in Table 4.6. The results showed that the empirical powers are close to nominal powers across all settings. As expected, power increased with increasing sample sizes, provided other parameter values were fixed. For example, the nominal and empirical powers both increased from 50% to around 70% when n is increased from 300 to 500 for $p_0 = 0.2, RR_x = 1.5$.

For cases involving two covariates, the simulation results adjusting for a binary covariate x_2 are presented in Table 4.7. The results suggest that the empirical power values are fairly closed to the nominal values, with a maximum difference of 5.4% observed at $p_0 = 0.3, RR_{x_1} = 1.5, RR_{x_2} = 1, r = 0.6, n = 300$. When $n = 300, p_0 = 0.2, RR_{x_2} = 1.5$, there is a more than 30% increase in both nominal power and empirical power by increasing RR_{x_1} from 1.5 to 2 for each r category. With fixed p_0, RR_{x_1} and RR_{x_2}, the nominal and empirical powers decrease when r is stronger. For instance, the nominal power decreases from 93.9% to 80.1% for $n = 500$, $RR_{x_1} = 1.5, RR_{x_2} = 2$ and the empirical power drops from 97% to 79.8% when r jumps from zero to strong for $p_0 = 0.2$.

Table 4.8 presents simulation results when the factor of interest is binary, and x_2 is continuous. The results show a similar power pattern as that for the two binary covariates scenario. The nominal and empirical power difference is less than 8.7%. Both powers drop when r goes up for the same parameter combination.

The simulation results from Tables 4.7 and 4.8 are also graphically presented in Figures 4.3 and 4.4 for the easy viewing purpose when $p_0 = 0.1$. Figures present that the nominal powers and empirical powers are similar to each other for detecting RR_{x_1}effect for all RR_{x_2} and r combinations. The figures display the similarity between the nominal power and empirical power. When x_2 is a continuous variable, the difference between the two powers is more obvious in Figure 4.4 compared to Figure 4.3, and the empirical power underestimates the nominal power slightly when the correlation is strong. The power trends for p_0 other than 0.1 are similar to the trend in Figures 4.3 and 4.4.

Table 4.6 Results of the nominal power from the proposed power formula ($NP_{proposed}$) for estimating RR, and the empirical power from the modified Poisson model ($EP_{mpoisson}$) for one binary covariate. The risk ratio $RR_x \in (1,1.5,2,2.5)$, the response probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. Empirical power is based on 1000 runs, $n \in (300,500)$.

		n = 300			n = 500		
p_0	RRx	p	$NP_{proposed}$	$EP_{mpoisson}$	p	$NP_{proposed}$	$EP_{mpoisson}$
0.1	1	0.10	5.0	4.1	0.10	5.0	3.7
	1.5	0.12	26.3	25.5	0.12	40.2	38.9
	2	0.15	71.2	68.7	0.15	90.2	88.5
	2.5	0.17	95.5	93.5	0.17	99.7	99.4
0.2	1	0.20	5.0	4.3	0.20	5.0	4.3
	1.5	0.25	52.6	49.7	0.25	74.4	72.3
	2	0.30	97.6	96.8	0.30	99.9	99.7
	2.5	0.35	100.0	100.0	0.35	100.0	100.0
0.3	1	0.30	5.0	3.7	0.30	5.0	3.7
	1.5	0.37	77.5	75.7	0.37	93.9	93.0
	2	0.45	100.0	100.0	0.45	100.0	100.0
	2.5	0.52	100.0	100.0	0.52	100.0	100.0
0.4	1	0.40	5.0	4.0	0.40	5.0	4.9
	1.5	0.50	93.9	94.3	0.50	99.5	99.2
	2	0.60	100.0	100.0	0.60	100.0	100.0
	2.5	0.70	100.0	100.0	0.70	100.0	100.0

Table 4.7 Results of the nominal power from the proposed power formula ($NP_{proposed}$) for estimating RR, and the empirical power from the modified Poisson model ($EP_{mpoisson}$) for two binary covariates. The risk ratios $RR_{x_1} \in (1,1.5,2,2.5)$, $RR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. Empirical power is based on 1000 runs, $n \in (300, 500)$.

				n = 300			*n* = 500		
p_0	$RRx1$	$RRx2$	*r*	*p*	$NPproposed$	$EPmpoisson$	*p*	$NPproposed$	$EPmpoisson$
0.1	1	1	0	0.10	5.0	4.7	0.10	5.0	4.9
	1.5	1	0	0.13	26.3	24.6	0.13	40.2	39.1
			0.3	0.13	24.3	23.7	0.13	37.1	37.3
			0.6	0.13	18.4	20.3	0.13	27.7	30.8
	1.5	1.5	0	0.16	32.6	32.9	0.16	49.5	51.6
			0.3	0.16	30.1	30.3	0.16	45.9	46.8
			0.6	0.16	22.6	22.2	0.16	34.4	35.3
	1.5	2	0	0.19	39.1	41.0	0.19	58.5	61.5
			0.3	0.19	36.2	36.6	0.19	54.5	55.9
			0.6	0.19	27.0	26.5	0.19	41.3	41.7
	2	1	0	0.15	71.2	69.7	0.15	90.2	90.6
			0.3	0.15	67.2	66.8	0.15	87.4	86.9
			0.6	0.15	52.2	55.7	0.15	74.0	74.2
	2	1.5	0	0.19	82.2	81.4	0.19	96.1	96.6
			0.3	0.19	78.5	77.7	0.19	94.4	94.6
			0.6	0.19	63.5	65.1	0.19	84.5	83.7
	2	2	0	0.23	89.8	90.9	0.23	98.7	98.5
			0.3	0.23	86.9	87.0	0.23	97.8	97.9
			0.6	0.23	73.4	73.6	0.23	91.6	91.7
	2.5	1	0	0.18	95.5	94.3	0.18	99.7	99.8
			0.3	0.18	93.6	92.1	0.18	99.4	99.2
			0.6	0.18	83.2	83.7	0.18	96.5	95.8
	2.5	1.5	0	0.22	98.7	98.2	0.22	100.0	100.0
			0.3	0.22	97.9	97.4	0.22	100.0	99.9
			0.6	0.22	91.9	91.3	0.22	99.1	99.0
	2.5	2	0	0.26	99.7	99.9	0.26	100.0	100.0
			0.3	0.26	99.5	99.3	0.26	100.0	100.0
			0.6	0.26	96.6	96.9	0.26	99.8	99.9
0.2	1	1	0	0.20	5.0	3.2	0.20	5.0	4.4
	1.5	1	0	0.25	52.6	52.3	0.25	74.4	75.5
			0.3	0.25	48.9	50.0	0.25	70.3	71.8
			0.6	0.25	36.7	40.2	0.25	55.2	58.1
	1.5	1.5	0	0.31	65.7	67.9	0.31	86.3	87.9
			0.3	0.31	61.6	63.2	0.31	82.9	83.8
			0.6	0.31	47.3	48.6	0.31	68.5	69.7
	1.5	2	0	0.38	77.5	80.2	0.38	93.9	97.0
			0.3	0.38	73.6	76.9	0.38	91.7	94.4
			0.6	0.38	58.5	59.3	0.38	80.1	79.8
	2	1	0	0.30	97.6	97.2	0.30	99.9	100.0
			0.3	0.30	96.3	96.4	0.30	99.8	99.9
			0.6	0.30	88.2	90.6	0.30	98.2	98.6
	2	1.5	0	0.38	99.6	99.7	0.38	100.0	100.0
			0.3	0.38	99.3	99.5	0.38	100.0	100.0

			0.6	0.38	96.1	96.9	0.38	99.8	99.7
	2	2	0	0.45	100.0	99.9	0.45	100.0	100.0
			0.3	0.45	99.9	99.9	0.45	100.0	100.0
			0.6	0.45	99.1	99.5	0.45	100.0	100.0
	2.5	1	0	0.35	100.0	100.0	0.35	100.0	100.0
			0.3	0.35	100.0	100.0	0.35	100.0	100.0
			0.6	0.35	99.7	99.5	0.35	100.0	100.0
	2.5	1.5	0	0.44	100.0	100.0	0.44	100.0	100.0
			0.3	0.44	100.0	100.0	0.44	100.0	100.0
			0.6	0.44	100.0	100.0	0.44	100.0	100.0
	2.5	2	0	0.53	100.0	100.0	0.53	100.0	100.0
			0.3	0.53	100.0	100.0	0.53	100.0	100.0
			0.6	0.53	100.0	100.0	0.53	100.0	100.0
0.3	1	1	0	0.30	5.0	4.9	0.30	5.0	5.2
	1.5	1	0	0.38	77.5	77.1	0.38	93.9	94.4
			0.3	0.38	73.6	74.1	0.38	91.7	92.2
			0.6	0.38	58.4	63.8	0.38	80.1	81.0
	1.5	1.5	0	0.47	90.9	91.6	0.47	98.9	99.2
			0.3	0.47	88.1	88.7	0.47	98.2	98.6
			0.6	0.47	75.0	77.4	0.47	92.5	93.0
	1.5	2	0	0.56	97.8	99.2	0.56	99.9	100.0
			0.3	0.56	96.7	98.3	0.56	99.8	100.0
			0.6	0.56	88.9	92.8	0.56	98.4	99.5
	2	1	0	0.45	100.0	99.9	0.45	100.0	100.0
			0.3	0.45	99.9	99.9	0.45	100.0	100.0
			0.6	0.45	99.1	99.2	0.45	100.0	100.0
	2	1.5	0	0.56	100.0	100.0	0.56	100.0	100.0
			0.3	0.56	100.0	100.0	0.56	100.0	100.0
			0.6	0.56	100.0	99.9	0.56	100.0	100.0
	2.5	1	0	0.53	100.0	100.0	0.53	100.0	100.0
			0.3	0.53	100.0	100.0	0.53	100.0	100.0
			0.6	0.53	100.0	100.0	0.53	100.0	100.0
0.4	1	1	0	0.40	5.0	4.3	0.40	5.0	4.2
	1.5	1	0	0.50	93.9	92.8	0.50	99.5	99.5
			0.3	0.50	91.7	92.1	0.50	99.1	99.5
			0.6	0.50	80.1	83.5	0.50	95.2	95.7
	1.5	1.5	0	0.63	99.5	99.4	0.63	100.0	100.0
			0.3	0.63	99.1	99.3	0.63	100.0	100.0
			0.6	0.63	95.2	96.5	0.63	99.7	99.8
	2	1	0	0.60	100.0	100.0	0.60	100.0	100.0
			0.3	0.60	100.0	100.0	0.60	100.0	100.0
			0.6	0.60	100.0	100.0	0.60	100.0	100.0
	2.5	1	0	0.70	100.0	100.0	0.70	100.0	100.0
			0.3	0.70	100.0	100.0	0.70	100.0	100.0
			0.6	0.70	100.0	100.0	0.70	100.0	100.0

Table 4.8 Results of the nominal power from the proposed power formula ($NP_{proposed}$) for estimating RR, and the empirical power from the modified Poisson model ($EP_{mpoisson}$) for one binary and one continuous covariates. The risk ratios $RR_{x_1} \in (1,1.5,2,2.5), RR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the response probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. Empirical power is based on 1000 runs, $n \in (300,500)$.

				n = 300			n = 500		
p_0	$RRx1$	$RRx2$	r	p	$NP_{proposed}$	$EP_{mpoisson}$	p	$NP_{proposed}$	$EP_{mpoisson}$
0.1	1	1	0	0.10	5.0	4.0	0.10	5.0	3.6
	1.5	1	0	0.12	26.0	25.7	0.12	39.8	38.9
			0.3	0.12	24.1	25.5	0.12	36.8	37.3
			0.6	0.12	18.3	20.5	0.12	27.5	28.2
	1.5	1.5	0	0.13	28.2	28.5	0.14	44.1	42.8
			0.3	0.14	26.4	24.7	0.14	41.5	38.4
			0.6	0.14	20.2	19.0	0.14	31.4	29.1
	1.5	2	0	0.16	32.7	37.4	0.16	51.4	56.5
			0.3	0.16	30.8	30.1	0.17	48.8	48.6
			0.6	0.16	23.5	19.8	0.17	37.4	30.7
	2	1	0	0.15	70.4	68.8	0.15	89.7	88.3
			0.3	0.15	66.4	66.4	0.15	86.8	86.0
			0.6	0.15	51.5	53.9	0.15	73.3	74.2
	2	1.5	0	0.16	74.3	75.2	0.16	93.0	94.1
			0.3	0.16	71.9	71.4	0.17	91.7	91.2
			0.6	0.17	58.4	54.4	0.17	81.4	76.5
	2	2	0	0.18	80.6	84.1	0.19	96.4	98.1
			0.3	0.19	78.3	79.7	0.20	95.6	95.1
			0.6	0.20	66.1	57.4	0.21	88.7	81.6
	2.5	1	0	0.17	95.0	93.3	0.17	99.7	99.4
			0.3	0.17	93.1	92.1	0.17	99.4	98.7
			0.6	0.17	82.3	81.7	0.17	96.2	94.6
	2.5	1.5	0	0.18	96.5	97.1	0.19	99.9	100.0
			0.3	0.19	95.9	96.2	0.20	99.8	99.6
			0.6	0.20	88.9	85.4	0.21	98.7	97.5
	2.5	2	0	0.20	97.6	97.0	0.21	100.0	100.0
			0.3	0.21	97.5	98.3	0.22	99.9	99.9
			0.6	0.23	93.1	88.5	0.23	99.5	98.4
0.2	1	1	0	0.20	5.0	4.6	0.20	5.0	4.5
	1.5	1	0	0.25	52.1	50.5	0.25	73.9	72.4
			0.3	0.25	48.4	48.9	0.25	69.8	69.0
			0.6	0.25	36.3	37.8	0.25	54.8	53.7
	1.5	1.5	0	0.27	56.7	58.7	0.28	79.9	82.6
			0.3	0.27	53.6	56.4	0.28	76.6	79.5
			0.6	0.28	41.2	40.1	0.28	62.7	56.9
	1.5	2	0	0.29	60.3	67.1	0.30	84.0	88.6
			0.3	0.29	56.4	61.8	0.30	80.2	88.4
			0.6	0.30	45.2	46.1	0.31	67.6	69.3
	2	1	0	0.30	97.3	97.2	0.30	99.9	99.7
			0.3	0.30	96.0	95.9	0.30	99.8	99.6
			0.6	0.30	87.5	88.2	0.30	98.0	98.2
	2	1.5	0	0.32	98.3	98.8	0.32	100.0	100.0

			0.3	0.32	97.5	98.0	0.33	99.9	100.0
			0.6	0.33	92.3	92.4	0.34	99.4	99.0
	2	2	0	0.33	98.8	99.5	0.34	100.0	100.0
			0.3	0.34	98.3	99.4	0.34	100.0	100.0
			0.6	0.32	91.9	97.6	0.33	99.2	100.0
	2.5	1	0	0.34	100.0	100.0	0.34	100.0	100.0
			0.3	0.34	100.0	100.0	0.34	100.0	100.0
			0.6	0.34	99.6	99.9	0.34	100.0	100.0
	2.5	1.5	0	0.36	100.0	100.0	0.36	100.0	100.0
			0.3	0.37	100.0	100.0	0.37	100.0	100.0
			0.6	0.38	99.9	99.8	0.39	100.0	100.0
	2.5	2	0	0.35	100.0	100.0	0.36	100.0	100.0
			0.3	0.34	100.0	100.0	0.34	100.0	100.0
			0.6	0.33	99.3	100.0	0.33	100.0	100.0
0.3	1	1	0	0.30	5.0	3.9	0.30	5.0	3.8
	1.5	1	0	0.37	76.9	76.2	0.37	93.6	93.1
			0.3	0.37	73.0	74.2	0.37	91.3	91.4
			0.6	0.37	57.8	58.1	0.37	79.5	79.8
	1.5	1.5	0	0.39	80.4	83.0	0.40	95.9	97.4
			0.3	0.40	77.5	84.0	0.40	94.4	96.3
			0.6	0.40	63.8	63.8	0.41	85.4	83.3
	1.5	2	0	0.40	81.9	88.3	0.41	96.3	98.7
			0.3	0.37	73.1	81.1	0.38	92.5	96.7
			0.6	0.39	61.4	64.6	0.38	81.7	87.4
	2	1	0	0.44	100.0	100.0	0.44	100.0	100.0
			0.3	0.44	99.9	100.0	0.44	100.0	100.0
			0.6	0.44	99.0	99.2	0.44	100.0	100.0
	2	1.5	0	0.46	100.0	100.0	0.46	100.0	100.0
			0.3	0.45	99.9	100.0	0.45	100.0	100.0
			0.6	0.44	98.9	99.8	0.44	100.0	100.0
	2.5	1	0	0.51	100.0	100.0	0.51	100.0	100.0
			0.3	0.51	100.0	100.0	0.51	100.0	100.0
			0.6	0.51	100.0	100.0	0.51	100.0	100.0
0.4	1	1	0	0.40	5.0	4.5	0.40	5.0	4.9
	1.5	1	0	0.50	93.5	94.4	0.50	99.4	99.2
			0.3	0.50	91.2	93.5	0.50	99.0	99.1
			0.6	0.50	79.4	82.7	0.50	94.9	96.1
	1.5	1.5	0	0.51	94.8	96.6	0.51	99.6	99.9
			0.3	0.50	91.9	96.6	0.51	99.2	99.7
			0.6	0.50	79.8	81.2	0.50	95.1	96.7
	2	1	0	0.59	100.0	100.0	0.59	100.0	100.0
			0.3	0.59	100.0	100.0	0.59	100.0	100.0
			0.6	0.59	100.0	100.0	0.59	100.0	100.0
	2.5	1	0	0.40	100.0	100.0	0.40	100.0	100.0
			0.3	0.40	100.0	100.0	0.40	100.0	100.0
			0.6	0.40	99.9	100.0	0.40	100.0	100.0

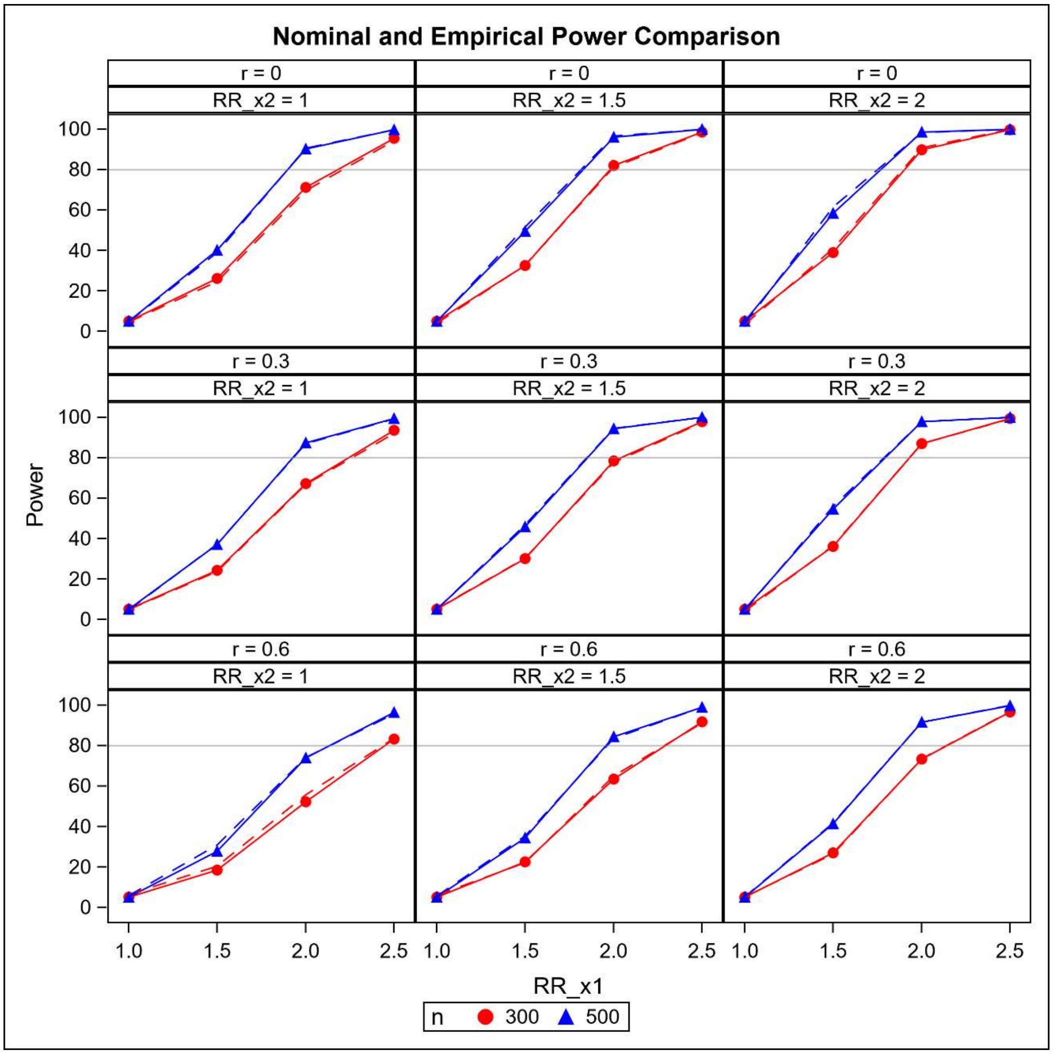

Figure 4.3 Comparison of nominal and empirical powers of testing for the effect of a binary risk factor x_1, adjusting for a binary covariate x_2 based on the simulation results from 1000 simulations of a sample size of 300 or 500 for $p_0 = 0.1$ as shown in Table 4.7. The RR_{x_1} and RR_{x_2} represent the risk ratios for x_1 and x_2, respectively, r represents the correlation between the two covariates, p_0 is the probability of outcome when both covariates are zero. Solid line represents the nominal power from Equation (3.10); dash line represents the empirical power of the modified Poisson model

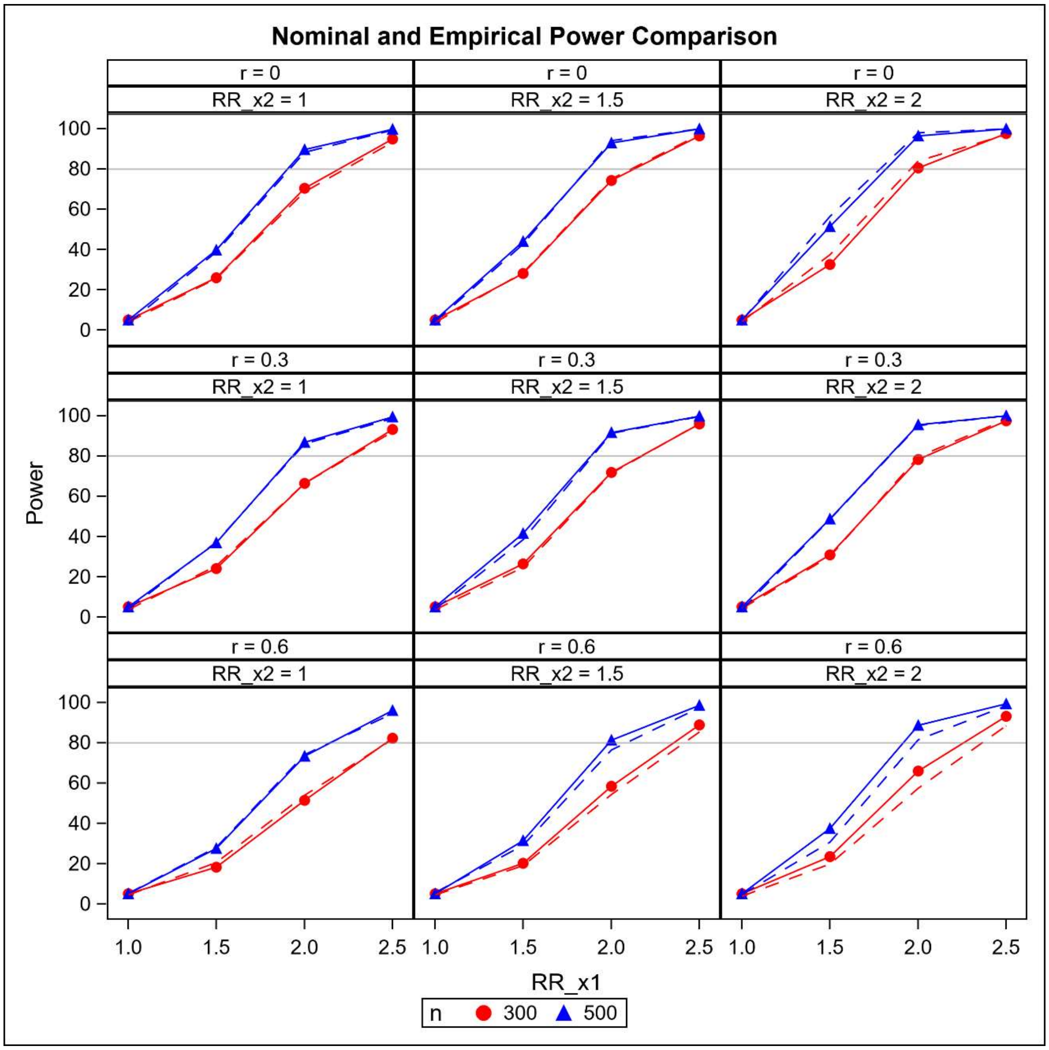

Figure 4.4 Comparison of nominal and empirical powers of testing for the effect of a binary risk factor x_1, adjusting for a continuous covariate x_2 based on the simulation results from 1000 simulations of a sample size of 300 or 500 for $p_0 = 0.1$ as shown in Table 4.8. The RR_{x_1} and RR_{x_2} represent the risk ratios for x_1 and x_2, respectively, r represents the correlation between the two covariates, p_0 is the probability of outcome when both covariates are zero. Solid line represents the nominal power from Equation (3.10); dash line represents the empirical power of the modified Poisson model

4.3.2.2. Power for a continuous risk factor

Tables 4.9 summarizes the simulation results for the powers of detecting the effect of a continuous risk factor, and Table 4.10 presents simulation results when adjusting for an additional continuous covariate. In both Tables, the powers reach over 95% when $RR_{x_1} = 1.5, n = 300$ for p_0 above 0.1. When n is 500 and $p_0 = 0.1$, $RR_{x_1} = 1.5$ brings both powers close to 90%. There are many settings with 100.0% powers in Tables 4.9 and 4.10, which means that the sample size used is larger than the sample size needed.

Compared to scenarios of two binary covariates and one binary one continuous covariates, the scenario with two continuous covariates shows a similar trend of power growth. Increasing the correlation still leads to a decrease in powers while fixing other parameter values. The largest difference between the nominal and empirical powers was observed at 7.7% when $p_0 = 0.1, RR_{x_1} = 1.5, RR_{x_2} = 2, r = 0.6, n = 300$.

The powers are higher in the two continuous covariates scenario than in the other two-covariate scenarios when $RR_{x_1} > 1$, considering the continuous interest may contain more information or have a larger variance than a binary interest. For example, if $p_0 = 0.3, RR_{x_1} = 1.5, RR_{x_2} = 2, r = 0.6, n = 300$, the nominal and empirical powers are respectively 97.4% and 97.3% when both covariates are continuous; 88.9% and 92.8% when both covariates are binary; and 61.4% and 64.6% when covariates are one binary and one continuous. If the factor of interest is a continuous variable and adjusting for another continuous variable, for obtaining an 80% power, a smaller sample size may be needed compared to other scenarios when $RR_{x_1} > 1$.

Table 4.9 Results of the nominal power from the proposed power formula ($NP_{proposed}$) for estimating RR, and the empirical power from the modified Poisson model ($EP_{mpoisson}$) for one continuous covariate. The risk ratio $RR_x \in (1,1.5,2,2.5)$, the probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. Empirical power is based on 1000 runs, $n \in (300,500)$.

		n = 300			n = 500		
p_0	RR_X	p	$NP_{proposed}$	$EP_{mpoisson}$	p	$NP_{proposed}$	$EP_{mpoisson}$
0.1	1	0.10	5.0	5.5	0.10	5.0	5.7
	1.5	0.11	69.7	72.5	0.11	88.6	88.0
	2	0.13	99.7	99.9	0.13	100.0	100.0
	2.5	0.15	100.0	100.0	0.14	100.0	100.0
0.2	1	0.20	5.0	5.2	0.20	5.0	6.0
	1.5	0.22	96.3	96.9	0.22	99.8	99.6
	2	0.25	100.0	100.0	0.24	100.0	100.0
	2.5	0.26	100.0	100.0	0.25	100.0	100.0
0.3	1	0.30	5.0	4.9	0.30	5.0	5.2
	1.5	0.33	99.8	99.8	0.32	100.0	100.0
	2	0.34	100.0	100.0	0.34	100.0	100.0
	2.5	0.35	100.0	100.0	0.34	100.0	100.0
0.4	1	0.40	5.0	5.2	0.40	5.0	5.2
	1.5	0.43	100.0	100.0	0.43	100.0	100.0
	2	0.44	100.0	100.0	0.43	100.0	100.0
	2.5	0.40	100.0	100.0	0.39	100.0	100.0

Table 4.10 Results of the nominal power from the proposed power formula ($NP_{proposed}$) for estimating RR, and the empirical power from the modified Poisson model ($EP_{mpoisson}$) for two continuous covariates. The risk ratios $RR_{x_1} \in (1,1.5,2,2.5), RR_{x_2} \in (1,1.5,2)$, the correlation $r \in (0,0.3,0.6)$, the probability of the baseline group $p_0 \in (0.1,0.2,0.3,0.4)$, and p represents the overall prevalence. Empirical power is based on 1000 runs, $n \in (300,500)$.

				n = 300			n = 500		
p_0	$RRx1$	$RRx2$	r	p	$NP_{proposed}$	$EP_{mpoisson}$	p	$NP_{proposed}$	$EP_{mpoisson}$
0.1	1	1	0	0.10	5.0	5.8	0.10	5.0	5.3
	1.5	1	0	0.11	68.7	69.8	0.11	88.6	87.2
			0.3	0.11	64.6	63.9	0.11	85.5	83.2
			0.6	0.11	49.9	46.9	0.11	71.6	68.4
	1.5	1.5	0	0.12	72.3	74.5	0.12	90.5	90.6
			0.3	0.12	70.7	74.4	0.12	89.5	90.6
			0.6	0.13	56.7	63.7	0.13	78.7	80.4
	1.5	2	0	0.13	78.9	84.6	0.13	94.2	95.9
			0.3	0.14	77.5	79.9	0.14	93.5	94.0
			0.6	0.15	63.9	71.6	0.14	84.3	86.0
	2	1	0	0.13	99.6	99.9	0.13	100.0	100.0
			0.3	0.13	99.2	99.5	0.13	100.0	100.0
			0.6	0.13	95.5	96.8	0.13	99.7	99.4
	2	1.5	0	0.13	99.7	99.5	0.13	100.0	100.0
			0.3	0.14	99.6	98.8	0.14	100.0	100.0
			0.6	0.14	97.7	97.0	0.14	99.9	99.8
	2	2	0	0.15	99.9	99.9	0.15	100.0	100.0
			0.3	0.15	99.8	98.5	0.15	100.0	100.0
			0.6	0.16	98.6	97.4	0.16	100.0	99.9
	2.5	1	0	0.14	100.0	100.0	0.14	100.0	100.0
			0.3	0.14	100.0	100.0	0.14	100.0	100.0
			0.6	0.14	99.9	99.3	0.14	100.0	100.0
	2.5	1.5	0	0.14	100.0	100.0	0.14	100.0	100.0
			0.3	0.15	100.0	99.9	0.15	100.0	100.0
			0.6	0.16	100.0	99.8	0.15	100.0	100.0
	2.5	2	0	0.15	100.0	100.0	0.15	100.0	100.0
			0.3	0.16	100.0	100.0	0.16	100.0	100.0
			0.6	0.17	100.0	100.0	0.17	100.0	100.0
0.2	1	1	0	0.20	5.0	5.1	0.20	5.0	5.6
	1.5	1	0	0.22	95.9	97.0	0.22	99.8	99.9
			0.3	0.22	94.1	95.3	0.22	99.5	99.7
			0.6	0.22	84.0	85.1	0.22	96.9	97.3
	1.5	1.5	0	0.23	97.0	95.7	0.23	99.9	99.8
			0.3	0.24	96.2	95.1	0.23	99.8	99.7
			0.6	0.24	88.5	86.6	0.24	98.3	97.9
	1.5	2	0	0.24	97.8	97.4	0.25	99.9	99.7
			0.3	0.25	97.1	97.4	0.24	99.8	99.8
			0.6	0.25	90.5	91.5	0.25	98.7	99.0
	2	1	0	0.24	100.0	100.0	0.24	100.0	100.0
			0.3	0.24	100.0	100.0	0.24	100.0	100.0
			0.6	0.24	100.0	100.0	0.24	100.0	100.0
	2	1.5	0	0.24	100.0	100.0	0.24	100.0	100.0
			0.3	0.25	100.0	100.0	0.25	100.0	100.0

			0.6	0.25	100.0	100.0	0.25	100.0	100.0
	2	2	0	0.25	100.0	99.9	0.24	100.0	100.0
			0.3	0.25	100.0	99.9	0.25	100.0	100.0
			0.6	0.24	100.0	99.9	0.24	100.0	100.0
	2.5	1	0	0.25	100.0	100.0	0.25	100.0	100.0
			0.3	0.25	100.0	100.0	0.25	100.0	100.0
			0.6	0.25	100.0	100.0	0.25	100.0	100.0
	2.5	1.5	0	0.26	100.0	100.0	0.25	100.0	100.0
			0.3	0.26	100.0	100.0	0.26	100.0	100.0
			0.6	0.26	100.0	100.0	0.25	100.0	100.0
	2.5	2	0	0.24	100.0	100.0	0.25	100.0	100.0
			0.3	0.25	100.0	100.0	0.24	100.0	100.0
			0.6	0.24	100.0	100.0	0.24	100.0	100.0
0.3	1	1	0	0.30	5.0	6.1	0.30	5.0	4.6
	1.5	1	0	0.32	99.8	99.5	0.32	100.0	100.0
			0.3	0.32	99.6	99.0	0.32	100.0	100.0
			0.6	0.32	97.1	95.5	0.32	99.9	99.7
	1.5	1.5	0	0.33	99.9	99.7	0.33	100.0	100.0
			0.3	0.33	99.7	99.0	0.33	100.0	100.0
			0.6	0.33	97.8	96.6	0.33	99.9	99.9
	1.5	2	0	0.32	99.8	99.5	0.32	100.0	100.0
			0.3	0.33	99.6	99.1	0.32	100.0	100.0
			0.6	0.33	97.4	97.3	0.32	99.9	100.0
	2	1	0	0.33	100.0	100.0	0.34	100.0	100.0
			0.3	0.33	100.0	100.0	0.34	100.0	100.0
			0.6	0.33	100.0	100.0	0.34	100.0	100.0
	2	1.5	0	0.34	100.0	100.0	0.34	100.0	100.0
			0.3	0.33	100.0	100.0	0.34	100.0	100.0
			0.6	0.32	100.0	100.0	0.32	100.0	100.0
	2.5	1	0	0.34	100.0	100.0	0.34	100.0	100.0
			0.3	0.34	100.0	100.0	0.34	100.0	100.0
			0.6	0.34	100.0	100.0	0.34	100.0	100.0
0.4	1	1	0	0.40	5.0	4.9	0.40	5.0	5.1
	1.5	1	0	0.42	100.0	100.0	0.43	100.0	100.0
			0.3	0.42	100.0	100.0	0.43	100.0	100.0
			0.6	0.42	99.8	99.6	0.43	100.0	100.0
	1.5	1.5	0	0.42	100.0	99.8	0.41	100.0	100.0
			0.3	0.41	100.0	99.9	0.41	100.0	100.0
			0.6	0.41	99.7	99.6	0.41	100.0	100.0
	2	1	0	0.42	100.0	100.0	0.42	100.0	100.0
			0.3	0.42	100.0	100.0	0.42	100.0	100.0
			0.6	0.42	100.0	100.0	0.42	100.0	100.0
	2.5	1	0	0.39	100.0	100.0	0.39	100.0	100.0
			0.3	0.39	100.0	100.0	0.39	100.0	100.0
			0.6	0.39	100.0	100.0	0.39	100.0	100.0

4.3.3. Conclusion

In this chapter, we evaluated the performance of the empirical powers from the modified Poisson model and the logistic regression, and we also assessed the proposed RR sample size and power formulas. When there is no effect for covariate x or x_1, the power of the hypothesis test should be at the significance level (Burton et al., 2006). This was validated from the simulation study that the empirical powers were all close to 5% under the scenarios of $\beta_1 = 0$ as shown in Tables 4.1 to 4.10.

When the underlying data set was generated using the logistic model, the simulation study showed that the empirical powers were similar between the modified Poisson model and the logistic regression, suggesting that the modified Poisson model can be applied in lieu of the logistic model in prospective studies with binary outcomes. The difference between the empirical powers from the two models was less than 5.6% in all scenarios considered in Section 4.3.1. However, when the correlation between two continuous covariates was large, the nominal power was higher than the empirical power for the logistic regression, and the difference could be as large as 14.3%. It indicates that having stronger correlation influences the actual power more when both covariates are continuous variables.

The simulation study assessed the performance of the proposed sample size formulas for the modified Poisson model. The proposed power formula performed well in all scenarios considered in Section 4.3.2. The empirical power closely agreed with the nominal power with fixed sample sizes. The power difference was less than 8.7% in all settings. The difference varied in scenarios with one or two continuous covariates. For example, the empirical power could be as much as 8.7% lower or 8.2% higher than the nominal power when $RR_{x_1} = RR_{x_2} = 2, r = 0.6, p_0 = 0.1, n = 300$ or $RR_{x_1} = 1.5, RR_{x_2} = 2, r = 0.3, p_0 = 0.2, n = 500$ in one binary and one continuous covariates scenario. These discrepancies could be due to the multicollinearity problem among covariates or simply due to sampling variations.

Chapter 5 Illustrating Examples

This chapter presents the application on the formula by using a subset of the Diabetes Control and Complications Trial (DCCT) database from Lachin (2011, p. 298). Zou (2004) also used the same data set for illustrative purposes. Section 5.1 introduces the DCCT database and the subset with descriptive statistics. Section 5.2 describes the specifications of the analysis and presents the results.

5.1. The data

Diabetes mellitus is a disease that relates to metabolic disorders and a high blood sugar level. The blood sugar (glucose) level is controlled by the hormone insulin produced from the pancreas. There are many possible symptoms of diabetes, including extreme hunger, fatigue, and weight loss.

The DCCT is a controlled randomization trial conducted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The DCCT study (1993; 2010) examined the influence of intensive insulin treatment on retinopathy and other complications among insulin-dependent diabetes mellitus (IDDM) patients. Retinopathy is a complication that may cause vision problems, and it leads to blindness if it becomes severe. A patient is more likely to develop such a complication when having a long diabetes history. The DCCT study started in August 1983 and ended in April 1993, and it enrolled 1441 patients, including the primary prevention cohort of 726 patients without retinopathy and the secondary intervention cohort of 715 patients with mild retinopathy. Each patient was randomly assigned to intensive insulin therapy. The average follow-up time was 6.5 years.

The DCCT research group (1993) provided basic summary statistics in their publication. The age of the patients ranged between 13 and 39 years. More than 95% of patients in each cohort were white. Besides retinopathy, the NIDDK studied the influence of intensive insulin therapy on other neurological, cardiovascular, and renal outcomes and found that the intensive insulin therapy was effective in delaying the progression of retinopathy and the onset of other complications.

Nephropathy (kidney disease) is another serious complication of diabetes. The DCCT Nephropathy (Microalbuminuria) subset from Lachin (2011) is used in the next section. The DCCT Nephropathy subset contains the records of 172 patients. The patients were in the secondary intervention cohort with baseline albumin excretion rates between 15 mg/24h and 40 mg/24h. Microalbuminuria ($micro24$) is the binary outcome that was evaluated at six years in the subset (Lachin, 2011, p. 298). The $micro24$ was assigned as 1 if the patient was diagnosed with microalbuminuria, otherwise $micro24 = 0$. In addition to the microalbuminuria ($micro24$), treatment group (int), the baseline HbA$_{1c}$ ($hbael$, glycated hemoglobin level), the prior duration of diabetes ($duration$) in months, the level of systolic blood pressure (sbp), and gender ($female$) are included in the subset. The int and $female$ are also binary and coded with 1 and 0. HbA$_{1c}$ is the average level of blood glucose control over the preceding 4 to 6 weeks before patients joined the trial. $yearsdm$ is an additional variable that converts the duration from months to years.

Table 5.1 presents simple descriptive statistics of the variables in the DCCT subset. In the subset, the difference in the number of patients between the intensive and conventional treatment groups is 6. Within the conventional therapy group, 37.35% of patients are diagnosed with microalbuminuria, and 12.36% are confirmed for microalbuminuria in the intensive therapy group. Each therapy group has slightly more male patients than female patients, and 28.72% of male patients and 19.23% of female patients have microalbuminuria.

Table 5.1 Descriptive statistics of DCCT Nephropathy subset.

Binary variables	n_{yes}	n_{no}	Mean	Var
Microalbuminuria	42	130	0.244	0.186
Intensive treatment	89	83	0.517	0.251
Female	78	94	0.453	0.249

Continuous variables	Mean	Var	Min	Max
HbA$_{1c}$	9.262	2.178	6.660	14.370
Duration(years)	9.430	11.149	1.333	15.000
Systolic blood pressure	116.326	116.829	90.000	148.000

5.2. Sample size calculation

In this section, we apply the proposed sample size formulas on the DCCT subset, focusing on the influence from intensive insulin treatment or HbA$_{1c}$. The treatment variable is considered as the factor of interest x_1 in Section 5.2.1, and HbA$_{1c}$ is treated as x_1 in Section 5.2.2. The intensive therapy group is used as a baseline. The derived OR sample size formula from the nominal power equation (Equation 4.1) is also applied for comparison.

5.2.1. Sample size for treatment effect

Supposing that we are interested in whether the intensive therapy could reduce the risk of microalbuminuria, Table 5.2 summarizes the relationship between the two variables. The RR and OR of microalbuminuria between the conventional and intensive treatment groups were $RR_{int} \approx 3.022$ (95% CI: 1.627, 5.614) and $OR_{int} \approx 4.227$ (95% CI:1.953, 9.150). The coefficient of determination of int on the other four possible risk factors was estimated to be 0.0008.

Table 5.2 2 x 2 table between Microalbuminuria and treatment therapy.

	Microalbuminuria (Yes)	Microalbuminuria (No)
Intensive treatment	11	78
Conventional treatment	31	52

Applying the logistic regression and the modified Poisson models to the subset and adjusting for $hbael, yearsdm, sbp, female$, the estimated regression coefficients of int were 1.583 (95% CI: 0.747, 2.420) and 1.080 (95% CI:0.484, 1.677), respectively. The intensive treatment group was used as the baseline group. The corresponding estimated OR and RR were 4.870 and 2.945. Identical results were also obtained by Zou (2004). The two estimated values were close to the OR_{int} and RR_{int} without adjustment. Both models confirmed that intensive insulin therapy could reduce the risk of getting microalbuminuria, as zero was not included in the confidence intervals and P-value < 0.05.

The proposed risk ratio sample size formula and the OR sample size equation from Vittinghoff et al. (2012) were utilized in finding n. We assumed that researchers could use either RR or OR sample size equation in study planning. As well as the unadjusted RR_{int} and OR_{int}, we assumed the mean of $micro24$, and the variance of int from Table 5.1 could represent the population parameter values. The subset came from a randomized clinical trial, so the coefficient of determination between int and other covariates was assumed to be zero. To achieve 80% power at $\alpha = 0.05$, the sample sizes required for RR and OR, respectively, are

$$\begin{aligned}
n_{RR} &= \frac{\left(Z_{1-\alpha/2} + Z_{\gamma}\right)^2}{(\beta_1^*)^2 \sigma_{x_1}^2} \frac{1-p}{p} \frac{1}{1 - r_{1,2\ldots k}^2} \\
&= \frac{(1.96 + 0.84)^2}{\left(ln(3.022)\right)^2 0.251} \frac{1 - 0.244}{0.244} \frac{1}{1 - 0} \\
&\approx 80, \\
n_{OR} &= \frac{\left(Z_{1-\alpha/2} + Z_{\gamma}\right)^2}{(\beta_1^*)^2 \sigma_{x_1}^2} \frac{1}{p(1-p)} \frac{1}{1 - r_{1,2\ldots k}^2} \\
&= \frac{(1.96 + 0.84)^2}{\left(ln(4.227)\right)^2 0.251} \frac{1}{(1 - 0.244)0.244} \frac{1}{1 - 0} \\
&\approx 82.
\end{aligned}$$

The n for a study estimating RR is 80, and the n for an OR study is 82. The two sample sizes are close to each other from the calculation.

Table 5.3 presents nominal powers for different sample sizes. With the same number of patients, the nominal power for a study with RR is almost same as that for the OR. Researchers need a sample size between 80 and 100 to obtain a power between 80% and 90% to detect the influence of intensive insulin treatment when using RR or OR.

Table 5.3 Nominal power (NP) for measuring the influence of intensive therapy with fixed sample sizes, effect measures are $RR_{int} = 3.022$ and $OR_{int} = 4.227, p = 0.244, \sigma_{int}^2 = 0.251, r^2 = 0$.

n_{int}	NP_{RR}	NP_{OR}
80	80.3	79.2
90	84.7	83.7

100	88.2	87.3

Varying the RR_{int} within the range of its confidence interval and having a fixed 80% power for the same population parameter values, the calculated sample sizes are presented in Table 5.4. When RR increases, the sample size decreases quickly.

Table 5.4 Sample size (n) for measuring intensive treatment effect with 80% power, $p = 0.244, \sigma_{int}^2 = 0.251, r^2 = 0, RR_{int}$ varies.

RR_{int}	n_{int}
2.0	202
3.0	81
4.0	51

With the sample size of 150, the minimum detectable effect for the treatment is 0.803 when the power is 80%, while keeping other parameter values unchanged. The minimum detectable effect of 0.803 corresponds to the risk ratio of 2.233, which represents the smallest RR to be detected with a sample size of 150 and an 80% of power under a two-tailed test.

5.2.2. Sample size for HbA$_{1c}$ effect

Consider that researchers wish to study the influence of HbA$_{1c}$ on microalbuminuria, we calculated the sample sizes for the continuous factor of interest HbA$_{1c}$, and this illustration is more related to an observational study. From Table 5.1, the mean of $micro24$ is 0.244, and the variance of HbA$_{1c}$ is 2.178. From the modified Poisson and logistic regression models applied on the DCCT subset, the unadjusted regression coefficients for the two models were 0.292 (95% CI:0.136, 0.447) and 0.433 (95 % CI: 0.187, 0.679) respectively. The corresponding effect measures are $RR_{\mathrm{HbA}_{1c}} = 1.339$ (95% CI: 1.146, 1.563) and $OR_{HbA_{1c}} = 1.542$ (95% CI: 1.206, 1.971). The coefficient of determination of HbA$_{1c}$ on other covariates was 0.066.

We assumed the mean of $micro24$, the variance of HbA_{1c}, the coefficient of determination between HbA_{1c} and other covariates, and the unadjusted regression coefficients for $RR_{\mathrm{HbA}_{1c}}$ and $OR_{\mathrm{HbA}_{1c}}$ from the subset can represent the population parameter values. At 80% power, the calculated sample sizes are

$$\begin{aligned} n_{RR} &= \frac{\left(Z_{1-\alpha/2}+Z_{\gamma}\right)^2}{(\beta_1^*)^2\sigma_{x_1}^2}\frac{1-p}{p}\frac{1}{1-r_{1,2\ldots k}^2} \\ &= \frac{(1.96+0.84)^2}{(0.292)^2 2.178}\frac{1-0.244}{0.244}\frac{1}{1-0.066} \\ &\approx 141, \\ n_{OR} &= \frac{\left(Z_{1-\alpha/2}+Z_{\gamma}\right)^2}{(\beta_1^*)^2\sigma_{x_1}^2}\frac{1}{p(1-p)}\frac{1}{1-r_{1,2\ldots k}^2} \\ &= \frac{(1.96+0.84)^2}{(0.433)^2 2.178}\frac{1}{(1-0.244)0.244}\frac{1}{1-0.066} \\ &\approx 112. \end{aligned}$$

There is a difference of 29 patients between the two sample sizes. If the multiple correlation coefficient is increased to 0.3, the sample size required would be 144 with the proposed RR sample size formula, and 115 with the OR sample size formula. We also calculated nominal powers for the sample sizes of 141, 160, and 180 in Table 5.5. Given the sample sizes, the test based on RR provided slightly less power than that based on OR, where the nominal power difference is less than 10%.

Table 5.5 Nominal power (NP) for measuring the influence of gender with fixed sample sizes, and effect measures are $RR_{Hb\quad_{1c}} = 1.339$ and $OR_{HbA_{1c}} = 1.542, p = 0.244, \sigma_{HbA_{1c}}^2 = 2.178, r^2 = 0.066$.

$n_{\mathrm{HbA}_{1c}}$	NP_{RR}	NP_{OR}
141	80.2	88.3
160	84.9	91.9
180	88.8	94.5

Varying the risk ratio of HbA_{1c} within the range of its confidence interval while having a fixed 80% power for the same assumed population parameter values for the mean of $micro24$, the variance of HbA_{1c}, and coefficient of determination, the calculated sample sizes are summarized in Table 5.6. The sample size needed is fewer when $RR_{\mathrm{HbA}_{1c}} = 1.5$ than when $RR_{\mathrm{HbA}_{1c}} = 1.2$.

Table 5.6 Sample size for measuring HbA_{1c} effect with 80% power, $\mathrm{p} = 0.244, \sigma^2_{HbA_{1c}} = 2.178, r^2 = 0.066, RR_{HbA_{1c}}$ varies.

$RR_{\mathrm{HbA}_{1c}}$	$n_{\mathrm{HbA}_{1c}}$
1.2	360
1.4	106
1.5	73

Using $n = 200$, the minimum detectable effect of hemoglobin level (HbA_{1c}) is 0.244, assuming the coefficient of determination was 0.066, variance of hemoglobin level was at 2.178, and the prevalence of macroalbuminuria was 0.244. It corresponds to the risk ratio of 1.277, which represents the smallest RR to be detected with a sample size of 200 and an 80% of power to reject the null hypothesis under a two-tailed test.

$$\begin{aligned}\beta^*_{HbA_{1c}} &= \frac{Z_{1-\alpha/2} + Z_\gamma}{\sigma_{x_1}\sqrt{n\frac{p}{1-p}\left(1 - r^2_{1,2\ldots k}\right)}} \\ &= \frac{1.96 + 0.84}{\sqrt{2.178}\sqrt{200\frac{0.244}{1-0.244}(1-0.066)}} \\ &= 0.244\end{aligned}$$

From the two examples, the application on the binary factor of interest showed that the calculated OR and RR sample sizes were close to each other. The application on the continuous factor of interest, on the other hand, produced distinctive OR and RR sample sizes. The application on the HbA_{1c} showed the sample sizes of detecting the effect when using OR and RR were different. This situation could belong to the simulation settings when the OR was 1.5 and correlation equaled 0, where the empirical power difference between the logistic regression and the modified Poisson model was greater than 5%, meanwhile the nominal powers of OR and RR were close to their empirical powers when the correlation was zero in Chapter 4.

Chapter 6 Discussion

The sample size determination is a critical question that researchers will face in planning a study. The existing literature describes many sample size methodologies for a study using the logistic regression to estimate OR. For a modified Poisson model to estimate RR, we derived the power and sample size formulas adjusting for a variance inflation factor for regression models with multiple covariates.

We first examined the power performance of the logistic regression and modified Poisson models via simulation studies. Our simulation study showed that the modified Poisson model provided equivalent power as the logistic model in the presence of multiple covariates in the model. We also evaluated the proposed power formula for the modified Poisson model by comparing their nominal power and empirical powers. Our simulation study showed that the proposed power formula performed well in most simulation settings, suggesting that the proposed power and sample size equations are adequate to estimate the number of subjects needed for perspective studies.

There are some limitations of the simulation study. First of all, the covariates in the simulation study were derived from the Bernoulli and standard normal distributions. We are uncertain about the power behavior of the proposed formula under other covariate distributions. Væth and Skovlund (2004) conducted a power analysis for the logistic and Cox regression models and showed that the nominal and empirical powers could still match when the Gamma distribution was applied in data-generating instead of a normal distribution. It can be anticipated that the empirical power would be close to the nominal power for the proposed sample size method if the Gamma distribution is implemented. Second, we did not consider negative regression coefficient values in the simulation. As the proposed RR sample size equation uses an absolute or a squared β_1^* value, having a negative RR in the calculation should not affect the sample size results.

With the DCCT nephropathy data, we provided usages of the RR sample size formula in practice. Depending on the purpose of a study, researchers should carefully select a sample size for a specific effect measure to obtain plausible research results.

The application on a binary or continuous factor of interest showed that the proposed sample size equation does not require an additional numerical approximation in the calculation. Researchers may merely plug-in preset parameter values as they need. The application also presented the proposed sample size equation can be implemented when multiple covariates are involved. Since many healthcare studies containing more than two covariates, the proposed equation is beneficial when constructing similar research studies.

We assumed binary outcomes in a study are independent in the derivations. Since clustered binary outcomes are quite common, arising either from studies with repeated measures or clinical trials randomizing intact social units instead of individuals. Thus, a new research topic is to develop sample size estimation methods for such situations in the context of the model proposed by Zou and Donner (2013).

Bibliography

Alam, M. K., Rao, M. B., & Cheng, F.-C. (2010). Sample Size Determination in Logistic Regression. *Sankhya B, 72*, 58–75.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj, 332*, 1080.

Bobbio, M., Demichelis, B., & Giustetto, G. (1994). Completeness of reporting trial results: effect on physicians' willingness to prescribe. *The Lancet, 343*, 1209–1211.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine, 25*, 4279–4292.

Cheung, Y. B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American journal of epidemiology, 166*, 1337–1344.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Routledge.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum Associates, Inc.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute, 11*, 1269–1275.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (Vol. 32). CRC press.

Demidenko, E. (2007). Sample Size Determination for Logistic Regression Revisited. *Statistics in medicine, 26*, 3385–3397.

Diabetes Control and Complications Trial Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England journal of medicine, 329*, 977–986.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials—a review. *Statistics in medicine, 3*, 199–214.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71*, 431–444.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology, 125*, 761–768.

Greenland, S., & Robins, J. M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics*, 55–68.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical science*, 29–46.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., & others. (1998). *Multivariate data analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in medicine, 8*, 795–802.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Statistics in medicine, 17*, 1623–1634.

Hsieh, F. Y., Lavori, P. W., Cohen, H. J., & Feussner, J. R. (2003). An Overview of Variance Inflation Factors for Sample-size Calculation. *Evaluation & the Health Professions, 26*, 239–257.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials, 2*, 93–113.

Lachin, J. M. (2011). *Biostatistical Methods: The Assessment of Relative Risks.* Wiley.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician, 36*, 158–160.

McNutt, L.-A., Wu, C., Xue, X., & Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American journal of epidemiology, 157*, 940–943.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*, pp. 370-384.

Neuhaus, J. M., & Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika, 80*, 807–815.

Nurminen, M. (1981). Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika, 68*, 525–530.

Nurminen, M. (1995). To use or not to use the odds ratio in epidemiologic analyses? *European journal of epidemiology, 11*, 365–371.

Ritz, J., & Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research, 13*, 309–323.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique*, 221–226.

Schwartz, L. M., Woloshin, S., & Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization*. Mass Medical Soc.

Self, S. G., & Mauritsen, R. H. (1988). Power/Sample Size Calculations for Generalized Linear Models. *Biometrics*, 79–86.

Sheps, M. C. (1958). Shall we count the living or the dead? *New England Journal of Medicine, 259*, 1210–1214.

Sinclair, J. C., & Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of clinical epidemiology, 47*, 881–889.

Spiegelman, D., & Hertzmark, E. (2005). Easy SAS Calculations for Risk or Prevalence Ratios and Differences. *American journal of epidemiology, 162*, 199–200.

U.S. National Library of Medicine. (2010, 3). Diabetes Control and Complications Trial (DCCT) - Full Text View - ClinicalTrials.gov. *Diabetes Control and Complications Trial (DCCT) - Full Text View - ClinicalTrials.gov*.

UCLA Institute for Digital Research & Education. (n.d.). SAS MACROS: CORR2DATA. *SAS MACROS: CORR2DATA*.

Vach, W. (2012). *Regression models as a tool in medical research.* CRC Press.

Væth, M., & Skovlund, E. (2004). A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine, 23*, 1781–1792.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models.* Springer New York.

Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *American journal of epidemiology, 123*, 174–184.

Walter, S. D. (2000). Choice of effect measure for epidemiological data. *Journal of clinical epidemiology, 53*, 931–939.

Walter, S. D., & Holford, T. R. (1978). Additive, multiplicative, and other models for disease risks. *American Journal of Epidemiology, 108*, 341–346.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817-838.

Whittemore, A. S. (1981). Sample Size for Logistic Regression with Small Response Probability. *Journal of the American Statistical Association, 76*, 27–32.

Wicklin, R. (2013). *Simulating Data with SAS.* SAS Institute. Retrieved from https://books.google.ca/books?id=PC7WboXY2W0C

Yelland, L. N., Salter, A. B., & Ryan, P. (2011). Performance of the Modified Poisson Regression Approach for Estimating Relative Risks from Clustered Prospective Data. *American journal of epidemiology, 174*, 984–992.

Zou, G. (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American journal of epidemiology, 159*, 702–706.

Zou, G. Y., & Donner, A. (2013). Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Statistical methods in medical research, 22*, 661–670.

Appendix: SAS coding

```
 /*one binary covariate, logit link*/;
options nocenter nonotes formdlim=' ';
/*output delivery system options on or off*/;
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
ODS TRACE ON;

%macro simdata(runs=, px=, b0=, b1=, n=);
 %ODSOFF;
/*generate data for one binary covariate*/
data covariates;
 call streaminit(123456); /*input random seeds*/;
   do i=1 to &n;
    pt=i;
     x = rand("bernoulli", &px);
       linpred = &b0 + &b1* x ;
       pi = logistic(linpred);
    output;
 end;
run;

/*calculate NP_logi using OR sample size equation(4.1)*/
data formula_logi;
overallp=logistic(&b0)*(1-&px)+logistic(&b0+&b1)*&px;
  inside = 1.96 - abs( &b1 )* sqrt(&px*(1-&px)) *
     sqrt( &n * overallp*(1- overallp)  );
  Power = (1 - probnorm(inside))*100;
  /*keep power;
run;

/*generate response with the same covariate for 1000 runs*/;
 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
       yi = rand("bernoulli", pi);
       keep run yi x pt;
     output;
end;
run;

proc sort; by run;
run;

/*apply the logistic regression on 1000 runs for EP_logi*/;
proc genmod data=sim desc;
     by run;
     model yi = x /dist=b link=logit;
     ods output ParameterEstimates=est_logi
```

```
         (where = (Parameter='x     ')  keep=run Parameter 
ProbChiSq );
run;

/*apply the modified Poisson model on 1000 runs for EP_mpo*/;
proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x/ dist=poisson link=log;
     repeated subject=pt/type=unstr;
     ods output GEEEmpPEst=est_mpo
         (where = (Parm='x')  keep=run Parm ProbZ );
run;


/*calculate the number of rejections of H_0 in 1000 runs*/;
data results_logi;
   set est_logi;
      rej = (ProbChiSq <0.05)*100;
run;


data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;
%ODSOn;

/*print NP_logi and EP_logi and EP_mpo results*/;
proc print data=formula_logi;
   var overallp power;
   title Formula Power_logi;
run;

proc means n mean data = results_logi;
  var rej;
  title Empirical Power_logi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;
%mend simdata;

/*input parameter values*/;
%simdata(runs= 1000, px=0.5, b0=-2.197, b1=0.693, n=300);


/*one continous covariate, logit link*/;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;

%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
```

```
%mend;
ODS TRACE ON;
%macro simdata(runs=, xmu=, xvar=,b0=, b1=,  n=);
 %ODSOFF;
data covariates;
  call streaminit(123456);
   do i=1 to &N;
    pt=i;
       x=rand('normal',&xmu,&xvar);
          linpred = &b0 + &b1* x ;
        pi = logistic(linpred);

    output;
 end;
run;

data pp;
  call streaminit(123456);
  set covariates;
     linpred = &b0 + &b1* x ;
        pi = logistic(linpred);
        yi = rand("bernoulli", pi);
        keep yi x;
 run;

ods output summary=avg_pi(keep= pi_mean );
proc means mean var data=covariates;var pi; run;

data formula_logi;
 set avg_pi;
  inside = 1.96 - abs( &b1 )* sqrt(&xvar) *
     sqrt( &n * pi_mean*(1- pi_mean) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;

        yi = rand("bernoulli", pi);
         keep run yi x pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
     model yi = x /dist=b;
     ods output ParameterEstimates=est_logi
         (where = (Parameter='x     ')  keep=run Parameter
ProbChiSq );
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x/ dist=poisson link=log;
     repeated subject=pt/type=unstr;
     ods output GEEEmpPEst=est_mpo

         (where = (Parm='x')  keep=run Parm ProbZ );
run;


data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;

data results_logi;
   set est_logi;
      rej = (ProbChiSq <0.05)*100;
run;


%ODSOn;

proc print data=formula_logi;
   var pi_mean power;
   title Formula Power_logi;
run;


proc means n mean data = results_logi;
  var rej;
  title Empirical Power_logi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;

%mend simdata;

%simdata(runs= 1000,xmu=0,xvar=1, b0=-2.197,b1=0.405,n=300);

/*2 binary covariates scenario, logit link*/;
/*The plug in SAS coding piece of RandMVBinary program can be
downloaded from
https://communities.sas.com/t5/SAS-IML-File-Exchange/Simulate-
Correlated-Multivariate-Binary-Variables/ta-p/221225*/;
/*the actual directory of the downloaded RandMVBinary program needs
to be input into %include "" in the following coding*/;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
%include "C:\...\RandMVBinary.sas";

%macro simdata(runs=, px1=, px2=,rho=,b0=, b1=, b2=,n=);
```

```
 %ODSOFF;
proc iml;
load module=_all_;
call randseed(123456);
/* from Wicklin (2013) _Simulating Data with SAS_, p. 157 */;
p = {&px1 &px2 };
R = { 1    &rho ,
     &rho   1   }; /* correlations */;
X = RandMVBinary(&n, p, R);
/* check results */
DiffMean = p - mean(X);
DiffCorr = R - corr(X);

create multix from X[colname={"x1" "x2" }];; /* create data set */;
append from X;       /* write data in vectors */;
close multix; /* close the data set */;
quit;

data covariates;
set multix;
linpred = &b0 + &b1* x1 +&b2*x2;;
pi = logistic(linpred);
pt=_N_;
run;

data formula_logi;
overallp=logistic(&b0)*(1-&px1)*(1-
&px2)+logistic(&b0+&b1+&b2)*&px1*&px2+logistic(&b0+&b1)*&px1*(1-
&px2)+logistic(&b0+&b2)*(1-&px1)*&px2;
  inside = 1.96 - abs( &b1 )* sqrt(&px1*(1-&px1)) *
     sqrt( &n * overallp*(1- overallp) *( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
     model yi = x1 x2 /dist=b;
     ods output ParameterEstimates=est_logi
         (where = (Parameter='x1     ')  keep=run Parameter
ProbChiSq );
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo
```

```
         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

data results_logi;
   set est_logi;
      rej = (ProbChiSq <0.05)*100;
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;

%ODSOn;

proc print data=formula_logi;
   var overallp power;
   title Formula Power_logi;
run;

proc means n mean data = results_logi;
  var rej;
  title Empirical Power_logi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;

%mend simdata;

%simdata(runs=1000, px1=0.5, px2=0.5,b0=-2.197, b1=0.405,
b2=0.405,rho=0.6,n=300);

/*check actual correlation between covariates if necessary, proc corr
data=covariates; run;*/;

/*1 binary 1 continuous, logit link*/;
/*The SAS program is revised based on the open source coding from
https://www.researchgate.net/post/Simulating-correlated-categorical-
and-continuous-variables-in-SAS*/;
/*the coding from the author Paul Anthony Dennis was based on
CORR2DATA from UCLA
https://stats.idre.ucla.edu/sas/sas/macros/sas-macros-corr2data/*/;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;

%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;

%macro simdata(runs=, px=,xmu=,xvar=,rho=,b0=, b1=, b2=,n=);

 %ODSOFF;
```

```
proc iml ;
 call randseed(123456);
C={1 &rho ,
    &rho 1 };
p = root(C); /*using a Cholesky transformation*/;
dim = nrow(C);
A = j(&n,1,.);
B = j(&n,1,.);
call randgen(A, 'BERNOULLI', &px);
call randgen(B, 'NORMAL',&xmu,&xvar);
myvar = A||B;
do i = 1 to dim;
myvar[, i] = myvar[,i]-(sum(myvar[,i])/&n);
end;
XX = (t(myvar)*myvar)/(&n-1);
U = root(inv(XX));
Y = myvar*T(U);
T = Y*p;
create outdata from T;
append from T;
quit;

data covariates;
set outdata;
if col1<0 then x1=0;
else if col1>0 then x1=1;
x2=col2;
drop col1-col2;
linpred = &b0 + &b1* x1 +&b2*x2;;
pi = logistic(linpred);
pt=_N_;
run;

ods output summary=avg_pi(keep=x1_var pi_mean );
proc means mean var data=covariates;var x1 pi; run;

data formula_logi;
set avg_pi;

 inside = 1.96 - abs( &b1 )* sqrt(&px*(1-&px)) *
     sqrt( &n * pi_mean*(1- pi_mean)*( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo
         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

proc genmod data=sim desc;
     by run;
     model yi = x1 x2 /dist=b;
     ods output ParameterEstimates=est_logi
         (where = (Parameter='x1     ')  keep=run Parameter
ProbChiSq );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;

data results_logi;
   set est_logi;
      rej = (ProbChiSq <0.05)*100;
run;

%ODSOn;

proc print data=formula_logi;
   var pi_mean power;
   title Formula Power_logi;
run;

proc means n mean data = results_logi;
  var rej;
  title Empirical Power_logi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;

%mend simdata;

%simdata(runs=1000, px=0.5,xmu=0,xvar=1,b0=-0.405, b1=0.405,
b2=0.693,rho=0.6,n=300);


/*two continuous covariates, logit link*/;
/*the program uses the RandNormal function in SAS iml */;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
ODS TRACE ON;
```

```
%macro simdata(runs=, xmu1=,xvar1=,xmu2=,xvar2=,rho=, b0=, b1=,b2=,
n=);
 %ODSOFF;

proc iml;
call randseed(123456);
Z={. . . .};
create covariates1 from Z[colname={'x1' 'x2' 'linpred' 'pi'}];
do i=1 to &n;

/* specify the mean and covariance of the population */;
Mean = {&xmu1, &xmu2};
R={1 &rho,
  &rho 1};  /*correlation matrix*/
sd1=sqrt(%SYSEVALF(&xvar1));
sd2=sqrt(%SYSEVALF(&xvar2));
SD=sd1||sd2;
Cov = Corr2Cov(R, sd); /* population covariances */;
X = RandNormal(1, Mean, Cov); /* generate covariates values*/;
x1=X[1,1];
x2=X[1,2];
linpred = &b0 + &b1* x1 +&b2*x2;
pi = logistic(linpred);
Z=X||linpred||pi;

append from Z;
end;

close;
quit;

data covariates;
set covariates1;
pt=_N_;
run;

ods output summary=avg_pi(keep=x1_var pi_mean );
proc means mean var data=covariates;var x1 pi; run;

data formula_logi;
set avg_pi;
 inside = 1.96 - abs( &b1 )* sqrt(&xvar1) *
     sqrt( &n * pi_mean*(1- pi_mean)*( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
     model yi = x1 x2 /dist=b;
     ods output ParameterEstimates=est_logi
         (where = (Parameter='x1     ')  keep=run Parameter
ProbChiSq );
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo

         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

/*rejection decision for H_0*/;
data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;
data results_logi;
   set est_logi;
      rej = (ProbChiSq <0.05)*100;
run;

%ODSOn;

proc print data=formula_logi;
   var pi_mean power;
   title Formula Power_logi;
run;
proc means n mean data = results_logi;
  var rej;
  title Empirical Power_logi;
run;
proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;

%mend simdata;

%simdata(runs=1000, xmu1=0,xvar1=1,xmu2=0,xvar2=1, b0=-2.197,
b1=0.405,b2=0.405, rho=0.3, n=300);


/*1 binary, anti log link*/;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
ODS TRACE ON;
```

```
%macro simdata(runs=, px=, b0=, b1=, n=);

 %ODSOFF;
data covariates;

 call streaminit(123456);
   do i=1 to &n;
    pt=i;
     x = rand("bernoulli", &px);
          linpred = &b0 + &b1* x ;
          pi = exp(linpred);
    output;
 end;
run;

data formula_mpo;
overallp=(1-&px)*exp(&b0)+&px*exp(&b0+&b1);
 inside = 1.96 - abs( &b1 )* sqrt(&px*(1-&px)) *
     sqrt( &n * overallp/(1- overallp) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x/ dist=poisson link=log;
     repeated subject=pt/type=unstr;
     ods output GEEEmpPEst=est_mpo

         (where = (Parm='x')  keep=run Parm ProbZ );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;

%ODSOn;

proc print data=formula_mpo;
   var overallp power;
   title Formula Power_mpo;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;
```

```
%mend simdata;

%simdata(runs= 1000, px=0.5, b0=-2.303, b1=0.405, n=500);

/*1 continuous, anti log link*/;

options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
ODS TRACE ON;

%macro simdata(runs=, xmu=, xvar=,b0=, b1=,  n=);

 %ODSOFF;
data covariates;
  call streaminit(123456);
   do i=1 to &N;
    pt=i;
     do until (pi<1);
       x=rand('normal',&xmu,&xvar);
          linpred = &b0 + &b1* x ;
          pi = exp(linpred);
    end;
    output;
 end;
run;

ods output summary=avg_pi (keep= pi_mean);
proc means mean var data=covariates;var pi; run;

data formula_avgpi;
set avg_pi;
 inside = 1.96 - abs( &b1 )* sqrt(&xvar) *
     sqrt( &n * pi_mean/(1- pi_mean) );
  Power = (1 - probnorm(inside))*100;

run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;

        yi = rand("bernoulli", pi);
         keep run yi x pt;
     output;
end;
run;

proc sort; by run;
run;
```

```
proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x/ dist=poisson link=log;
     repeated subject=pt/type=unstr;
     ods output GEEEmpPEst=est_mpo
         (where = (Parm='x')  keep=run Parm ProbZ );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;

%ODSOn;

proc print data=formula_avgpi;
   var pi_mean power;
   title Formula Power_avgpi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;

%mend simdata;
%simdata(runs= 1000,xmu=0,xvar=1, b0=-2.303,b1=0.405,  n=300);


/*2 binary covariates scenario, anti log link*/;
/*The plug in SAS coding piece of RandMVBinary program can be
downloaded from
https://communities.sas.com/t5/SAS-IML-File-Exchange/Simulate-
Correlated-Multivariate-Binary-Variables/ta-p/221225*/;
/*the actual directory of the downloaded RandMVBinary program needs
to be input into %include "" in the following coding*/;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
%include "C:\...\RandMVBinary.sas";

%macro simdata(runs=, px1=, px2=,rho=,b0=, b1=, b2=,n=);
 %ODSOFF;
proc iml;
load module=_all_;
call randseed(123456);
p = {&px1 &px2 };
R = { 1    &rho ,
     &rho   1   }; /* correlations */;
X = RandMVBinary(&n, p, R);

DiffMean = p - mean(X);
```

```
DiffCorr = R - corr(X);

create multix from X[colname={"x1" "x2" }];
append from X;
close multix;
quit;

data covariates;
set multix;
linpred = &b0 + &b1* x1 +&b2*x2;;
pi = exp(linpred);
pt=_N_;
run;

data formula_mpo;
overallp=exp(&b0)*(1-&px1)*(1-
&px2)+exp(&b0+&b1+&b2)*&px1*&px2+exp(&b0+&b1)*&px1*(1-
&px2)+exp(&b0+&b2)*(1-&px1)*&px2;
 inside = 1.96 - abs( &b1 )* sqrt(&px1*(1-&px1)) *
     sqrt( &n * overallp/(1- overallp)*( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo
         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;
%ODSOn;

proc print data=formula_mpo;
   var overallp power;
   title Formula formula_mp;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;
```

```
%mend simdata;

%simdata(runs=1000, px1=0.5, px2=0.5,b0=-2.303, b1=0.405,
b2=0.405,rho=0.6,n=300);

/*1 binary 1 continous, anti log link*/;
/*The SAS program is revised based on the open source coding from
https://www.researchgate.net/post/Simulating-correlated-categorical-
and-continuous-variables-in-SAS*/;
/*the coding from the author Paul Anthony Dennis was based on
CORR2DATA from UCLA
https://stats.idre.ucla.edu/sas/sas/macros/sas-macros-corr2data/*/;

options nocenter nonotes formdlim=' '; * mprint;
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;

%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;

%macro simdata(runs=, px1=,xmu2=,xvar2=,rho=,b0=, b1=, b2=,n=);

 %ODSOFF;

proc iml ;
 call randseed(123456);
C={1 &rho ,
    &rho 1 };
p = root(C);
dim = nrow(C);
A = j(10000,1,.); /*generate 10k for filtering pi>1 later*/;
B = j(10000,1,.);
call randgen(A, 'BERNOULLI', &px1);
call randgen(B, 'NORMAL',&xmu2,&xvar2);
myvar = A||B;
do i = 1 to dim;
myvar[, i] = myvar[,i]-(sum(myvar[,i])/10000);
end;
XX = (t(myvar)*myvar)/(10000-1);
U = root(inv(XX));
Y = myvar*T(U);
T = Y*p;
create outdata from T;
append from T;
quit;

data covariates1;
set outdata;
if col1<0 then x1=0;
else if col1>0 then x1=1;
x2=col2;
drop col1-col2;
linpred = &b0 + &b1* x1 +&b2*x2;;
pi = exp(linpred);
pt=_N_;
```

```
run;

data covariates2;
set covariates1;
if pi<1;
run;

data covariates;
set covariates2 (obs=&n);
run;

ods output summary=avg_pi(keep=x1_var pi_mean );
proc means mean var data=covariates;var x1 pi; run;

data formula_avgpi;
set avg_pi;

 inside = 1.96 - abs( &b1 )* sqrt(&px1*(1-&px1)) *
     sqrt( &n * pi_mean/(1- pi_mean)*( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc reg data=sim;
    model x1= x2;
       ods output  FitStatistics = bb  ( where=(Label2='R-Square')
                                      keep= Label2 nValue2
                                      rename=(nValue2=Rsquare));
quit;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo
         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;
%ODSOn;

proc print data=formula_avgpi;
   var pi_mean power;
   title Formula Power_avgpi;
run;
```

```
proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;
%mend simdata;

%simdata(runs=1000, px1=0.5,xmu2=0,xvar2=1, b0=-1.609,
b1=0.405,b2=0.405, rho=0.6, n=300);

/*two continuous covariates, anti log link*/;
/*the program uses the RandNormal function in SAS iml */;
options nocenter nonotes formdlim=' ';
%macro ODSOff();
ods graphics off;
ods exclude all;
ods noresults;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
ODS TRACE ON;

%macro simdata(runs=, xmu1=,xvar1=,xmu2=,xvar2=,rho=, b0=, b1=,b2=,
n=);
 %ODSOFF;
proc iml;
call randseed(123456);
Z={. . . .};
create covariates1 from Z[colname={'x1' 'x2' 'linpred' 'pi'}];
do i=1 to &n;
do until (pi<1);
/* specify the mean and covariance of the population */
Mean = {&xmu1, &xmu2};
R={1 &rho,
  &rho 1};  /*correlation matrix*/
sd1=sqrt(%SYSEVALF(&xvar1));
sd2=sqrt(%SYSEVALF(&xvar2));
SD=sd1||sd2;
Cov = Corr2Cov(R, sd); /* population covariances */
X = RandNormal(1, Mean, Cov);
x1=X[1,1];
x2=X[1,2];
linpred = &b0 + &b1* x1 +&b2*x2;
pi = exp(linpred);
Z=X||linpred||pi;
end;

append from Z;
end;

close;
quit;

data covariates;
set covariates1;
pt=_N_;
run;
```

```
ods output summary=avg_pi(keep=x1_var pi_mean );
proc means mean var data=covariates;var x1 pi; run;

data formula_avgpi;
set avg_pi;

 inside = 1.96 - abs( &b1 )* sqrt(&xvar1) *
     sqrt( &n * pi_mean/(1- pi_mean)*( 1- &rho**2) );
  Power = (1 - probnorm(inside))*100;
run;

 data sim;
    call streaminit(123456);
     set covariates;
     do run=1 to &runs;
        yi = rand("bernoulli", pi);
         keep run yi x1 x2 pt;
     output;
end;
run;

proc sort; by run;
run;

proc reg data=sim;
    model x1= x2;
       ods output  FitStatistics = bb  ( where=(Label2='R-Square')
                                      keep= Label2 nValue2
                                      rename=(nValue2=Rsquare));
quit;

proc genmod data=sim desc;
     by run;
       class pt;
     model yi = x1 x2/ dist=poisson link=log;
     repeated subject=pt/type=unstr ;
     ods output GEEEmpPEst=est_mpo

         (where = (Parm='x1')  keep=run Parm ProbZ );
run;

data results_mpo;
   set est_mpo;
      rej = (ProbZ <0.05)*100;
run;
%ODSOn;

proc print data=formula_avgpi;
   var pi_mean power;
   title Formula Power_avgpi;
run;

proc means n mean data = results_mpo;
  var rej;
  title Empirical Power_mpo;
run;
%mend simdata;
%simdata(runs=1000, xmu1=0,xvar1=1,xmu2=0,xvar2=1, b0=-2.303,
b1=0.405,b2=0.405, rho=0.6, n=500);
```

Curriculum Vitae

Name:	Zhenni Xue
Post-secondary Education and Degrees:	Iowa State University Ames, Iowa, USA 2009-2012 B.Sc.
	Stevens Institute of Technology Hoboken, New Jersey, USA 2013-2015 M.Sc.
	The University of Western Ontario London, Ontario, Canada 2018-2021 M.Sc.