



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

A Comparison of Estimation Methods for Relative Risk in Binary response



The Graduate School

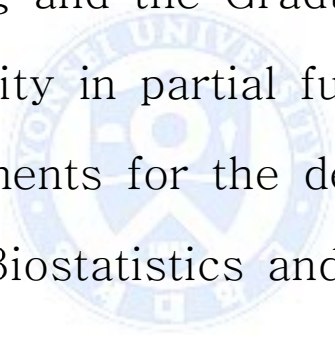
Yonsei University

Department of Biostatistics and Computing

A Comparison of Estimation Methods for Relative Risk in Binary response

A Master's Thesis

Submitted to the Department of Biostatistics
and Computing and the Graduate School of
Yonsei University in partial fulfillment of the
requirements for the degree of
Master of Biostatistics and Computing



Jihyun Lee

December 2015

This certifies that the master's thesis of
Jihyun Lee is approved.

남 정 모

Thesis Supervisor: Dr. Chung Mo Nam

정 인 경

Dr. Inkyung Jung

박 소 희

Dr. Sohee Park

The Graduate School
Yonsei University
December 2015

Acknowledgement

처음 대학원 생활을 시작 할 때는 이 과정의 끝이 보이지 않을 만큼 멀게 느껴졌는데 어느덧 시간이 지나 석사과정을 마치는 때가 되었습니다. 대학원 생활은 많은 일들이 있었고 많은 생각을 할 수 있었던 시간이었던 것 같습니다. 지금 이 순간이 오기까지 많은 분들이 있었기에 제가 논문을 완성할 수 있었던 것 같습니다. 제게 도움을 주신 분들께 감사의 마음을 전합니다.

늦은 나이의 시작으로 많은 어려움이 있었지만 마지막까지 논문을 마칠 수 있도록 신경 써주시고 지켜봐주신 남정모 교수님, 바쁘신 와중에도 지도해주시고 도움을 주신 교수님께 감사드립니다. 훌륭한 학업적인 가르침 이외에도 학생들을 위하는 마음으로 많은 말씀을 해주신 정인경 교수님, 송기준 교수님께 감사드립니다. 다양한 프로젝트를 통하여 값진 경험을 할수있도록 제게 기회를 주신 박소희 교수님 감사드립니다.

대학원 생활에서 가장 힘이되고 의지가 되었던 하나뿐인 제 동기 효진이에게 고맙다는 말을 전합니다. 저를 잘 챙겨주고 생각해주었던 성환선배, 성준씨, 늘 친절한 미소로 대해주던 승균씨, 함께한 시간이 적어 아쉬움이 있지만 가끔 만나더라도 친절히 대하였던 세휘, 세영이, 저를 잘 따라주고 스스로없이 도와주었던 세정이,지유,소정이에게 고맙다는 말을 전합니다.

보건대학원에서 연구원 생활을 하면서 진심어린 충고와 고민을 들어준던 정수연 선생님, 동고도락 하며 힘들 때 큰 힘이 되어주고, 마음을 토닥거리며 위로 해주었던 심성근 선생님, 장보원 선생님, 짧은 시간이었지만 제가 부탁할 때 도움주시고 마음써준 선지유 선생님에게 감사합니다. 사소한 일부터 큰일까지 자기 일처럼 신경써주시고 좋은 일, 힘든 일 있을 때마다 늘 함께해주던 큰 버팀목이었던 이예슬 선생님, 이미경 선생님 진심으로 고맙습니다.

소중한 가족에게 감사한 마음 전합니다. 늘 아낌없는 응원과 지원을 해주시는 사랑하는 부모님, 공부하는 저를 배려해주시고, 항상 걱정해주시던 시부모

님 감사합니다. 마지막으로 항상 저를 먼저 위하고 생각해주는 사랑하는 남편
에게 감사합니다.



Contents

List of Tables	ii
Abstract	iii
I . Introduction	1
II. Theoretical Background	3
2.1 Notations	3
2.2 Odds ratio and Relative risk	4
2.3 Logistic regression	4
2.4 Log-binomial regression	5
2.5 Poisson regression and Modified Poisson regression	6
III. Simulation	8
3.1 Log-binomial model	8
3.2 Binomial model	9
3.2 Simulation Result	10
3.2.1 Log-binomial model	10
3.2.2 Binomial model	13
IV. Illustrative data	15
V. Conclusion and Discussion	18
Bibliography	20
국문 요약	22

List of Tables

Table 1. Simulation results: Log-binomial Model, $n=1000$	12
Table 2. Simulation results: Binomial Model, $n=1000$	14
Table 3. Summary statistic for illustrative data.....	16
Table 4. Result of the estimated odds ratio and relative risk by different regression model.....	17



Abstract

The odds ratio and relative risk are usually the indices of interest in public health and medical studies. The odds ratio can be obtained using logistic regression in case-control studies. In cohort studies, however, the odds ratio should not be replaced with relative risk. This can cause overestimation or underestimation of the treatment effect in the study under some conditions. In this paper, we compare multiple methods to estimate the appropriate relative risk in a binary response. The odds ratio can be obtained using logistic regression. With an incidence of the outcome of more than 10%, the odds ratio should not be replaced with the relative risk. Log-binomial regression has become an alternative to logistic regression for the analysis. However, it fails to converge at a high incidence. The Poisson regression using a sandwich variance estimator outperforms in estimating the relative risk directly in terms of MLEs and the convergence problem. It is reliable in terms of simulation results. Data from a diabetes study are used to illustrate the different methods.

KEY WORDS: Odds ratio, Relative risk, Logistic regression, Log-binomial regression, Poisson regression, Modified Poisson regression, Log-binomial model, Estimating relative risk

I . Introduction

Odds ratios and relative risk are widely used to estimate risk in one group compared with another group in clinical trials and the public health field. In a case-control study, the odds ratio could be obtained directly using logistic regression. The odds ratio reflects relative risk, which is typically overestimated. Under some conditions, such as when the incidence of the outcome is less than 10%, it is acceptable to apply relative risk instead of the odds ratio (Zhang and Kai 1998). However, using the odds ratio exaggerates a treatment effect or risk association by more than 10% of the incidence of the outcome (Zhang and Kai 1998). This overestimation increases with increasing incidence (Knol et al. 2012).

There are alternative methods to estimate relative risk, such as log-binomial, Poisson, and modified Poisson regression (also called Poisson regression with robust standard errors) analyses. Log-binomial regression is a useful approach to estimate the correct risk ratio and associated confidence intervals. As for logistic regression, log-binomial regression is a generalized linear model (GLM), used to analyze a dichotomous outcome. The difference between log-binomial and logistic regression analyses is the link function. In log-binomial regression, a log link is used, but for logistic regression a logit link is used. Poisson regression is also a GLM, with a log link, and the dependent variable follows a Poisson distribution. Both the log-binomial and Poisson regression analyses are capable of estimating relative risk. However,

log-binomial regression could have problems with convergence. Standard errors obtained from Poisson regression analysis are typically large. Thus, the Poisson regression with a robust error variance could decrease the standard error and accurately estimate the relative risk and confidence intervals.

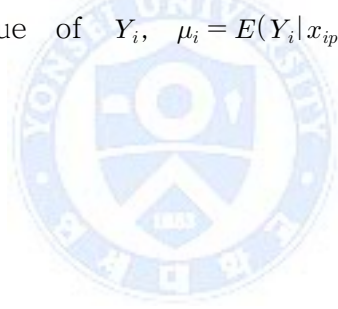
The purpose of this paper was to compare multiple methods to estimate adjusted relative risk. The methods were applied to the log-binomial and binomial models through simulation, under different conditions, such as changing incidence and strengthened exposure effect. The estimated relative risk, standard deviation, means of standard errors, and coverage rates were then compared. These methods were applied to a typical cohort study.

A summary section provides background information and the purpose of this study. Descriptions of the theoretical background, including the odds ratio, relative risk, and logistic, log-binomial, and Poisson regression analyses, with variance estimates, are provided in Section 2. Section 3 presents results from a simulation study and compares the methods used to estimate relative risk. The methods were applied to real cohort data; relative risk estimates are provided in Section 4. Finally, the discussion and conclusions are provided in Section 5.

II. Theoretical Background

2.1 Notations

In this study, we considered GLMs to estimate relative risk. In the GLM, there are three components required to specify the model. The random component identifies the response variable Y_i and follows a specified probability distribution. The systematic component represents the explanatory variables and follows a probability distribution $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. The third component, the link function, is $g(\cdot)$. The mean of expected value of Y_i , $\mu_i = E(Y_i | x_{ip})$, is specified by link function.



2.2 Odds ratio and relative risk

Comparing two groups on a binary response, Y, the data could be displayed in a contingency table. From the 2×2 contingency table, the measurement index of association, the odds ratio, and relative risk could be obtained.

Odds ratios represents a ratio of two odds,

$$OR = \frac{p_1 / (1 - p_1)}{p_0 / (1 - p_0)}$$

where p_0 is the probability of the outcome for the unexposed and p_1 is the probability of the outcome for the exposed. In other words, the odds ratio is the probability that an outcome occurs given an exposure, compared to the odds of the outcome occurring for a non-exposure. Whereas the odds ratio is a ratio of two odds, the relative risk is the ratio of two probabilities, defined as follows:

$$RR = \frac{p_1}{p_0}$$

2.3 Logistic regression

GLM that uses the logit link is called a logistic regression model and is widely used in modeling binary response variables. The model is expressed as

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}. \quad (1)$$

From Eq. (1), the regression coefficient represents differences in the log odds, $\exp(\beta_i) = OR_i$ for a one-unit increase in x_{i1} adjusted for all the other covariates.

2.4 Log-binomial regression

Log-binomial regression is similar to logistic regression, except for the link function. The log-binomial uses a log link function, rather than a logit function.

$$g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}. \quad (2)$$

In Eq. (2), β_i 's are differences in log risks so $\exp(\beta_i) = RR$ for a one-unit increase in x_{i1} adjusted for all the other variables.

2.5 Poisson and modified Poisson regression

The Poisson distribution is used as a discrete distribution to model

count data. This distribution is unique, in that its mean and variance are equivalent (Hosseinian 2009). So we take the logarithm and apply the following model.

$$g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \quad (3)$$

This is a classical regression model. However, if the Poisson mean is related to regressors x_{i1}, \dots, x_{ip} , as in Eq. (3), then the variance is

$$\text{var}(Y_i) = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}.$$

It is shown that the variance depends on the regressors and so the equal variance assumption is not accounted. However, the Poisson distribution assumes that the sum of independent Poisson random variables is Poisson as well. (Winkelmann 2013) Therefore, we can have a log-linear Poisson model. Poisson has to non-negative value, we should take logarithm.

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \dots, n. \quad (4)$$

In order to estimate the parameters of Poisson regression model, maximum likelihood estimation is commonly used. The log-likelihood

function is

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^n (y_i \log \mu_i - \mu_i). \quad (5)$$

To satisfy the goal that finding the values of β that maximize the Eq.(5), Eq.(5) is differentiated with respect to β , $\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^n \mathbf{x}_i (y_i - \mu_i)$ and set the result to zero. (Hosseinian 2009) Thus, application of this estimation equation results in consistent estimators, as given by the solution to the score equation provided below (Zou and Donner 2013) :

$$\sum_{i=1}^n \mathbf{x}_i (y_i - \mu_i(\beta)) = 0 \quad (6)$$

Use of the Poisson model for binary data shows an inaccurately specified variance function. Therefore, using a sandwich variance estimator, the variance estimator for $\hat{\beta}$ is

$$\widehat{var}(\hat{\beta}) = A^{-1}BA \quad (7)$$

where $A = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' e^{\mathbf{x}_i' \beta}$ and $B = [\sum_{i=1}^n \mathbf{x}_i (y_i - e^{\mathbf{x}_i' \beta})][\sum_{i=1}^n (y_i - e^{\mathbf{x}_i' \beta}) \mathbf{x}_i']$.

III. Simulation

In this section we conducted simulations to evaluate the performance of the log-binomial and binomial models, under different scenarios. The simulated dataset included a dichotomous exposure X , a dichotomous exposure Y , and five dichotomous confounders $\mathbf{Z} = (Z_1, Z_2, \dots, Z_5)$. We compared the mean estimates (based on 1,000 replicates), the empirical standard deviations of the parameter estimates, the mean values of the estimated standard errors, and the coverage probability of the 95% confidence interval.

3.1 Log-binomial model

The true response model was assumed as a log-binomial model.

$$E(Y|X, \mathbf{Z}) = \exp(\beta_0 + \beta_1(X - 0.5) + \sum_{i=1}^5 \gamma_i(Z_i - 0.5)) \quad (8)$$

The confounders (\mathbf{Z}) were independent and dichotomous, with 50% incidence, and they generated a binomial distribution with a probability of 0.5. The exposure X was generated from a binomial distribution, with a success probability of $\Pr(X=1|\mathbf{Z}) = 0.5 * \exp(\alpha(\mathbf{Z} - 0.5))$, where the parameter $\alpha = \log(1.13)$ represents the effect of the confounder \mathbf{Z} on the exposure \mathbf{X} . In the simulation study, baseline incidence and the exposure effect were changed. The baseline incidence outcome is $\beta_0 = \log(\text{incidence})$. At first, we set the baseline incidence at 5% and

changed up to 40% by 5%.The exposure effect $\exp(\beta_1)$ is 0.7,1.5,and 3.0. (Knol et al. 2012)

3.2 Binomial model

True response model was assumed as binomial model.

$$E(Y|X, \mathbf{Z}) = \beta_0 + \beta_1 X + \sum_{i=1}^5 \gamma_i (Z_i - 0.5) \quad (9)$$

The confounders (\mathbf{Z}) were independent and dichotomous. They were generated from a binomial distribution, with a probability of 0.5. The exposure, X , was generated from a binomial distribution with a success probability of $\Pr(X=1|\mathbf{Z}) = \alpha + 0.05 \sum_{i=1}^5 Z_i$. The parameter, α , indicates the proportion exposed, which was 50%. In the simulation, baseline incidence and the exposure effect varied. the baseline incidence β_0 (5-40%, increasing by increments of 5%) and the exposure effect β_1 (0.7,1.5,and 3.0). In this model, the exposure effect (β_1) was not relative risk; relative risk is estimated as follows.

Let's define probability of $X=0$ and $X=1$ as follows :

$$\Pr(Y=1|X=0, \mathbf{Z}) = \beta_0 + \sum_{i=1}^5 \gamma_i Z_i \quad (10)$$

$$\Pr(Y=1|X=1, \mathbf{Z}) = \beta_0 + \beta_1 + \sum_{i=1}^5 \gamma_i Z_i. \quad (11)$$

Here, the relative risk can be expressed by

$$RR = \frac{P(Y=1|X=1, \mathbf{Z})}{P(Y=1|X=0, \mathbf{Z})} = 1 + \frac{\beta_1}{P(Y=1|X=0, \mathbf{Z})} . \quad (12)$$

The relative risk can vary with the value of \mathbf{Z} . The regression coefficient β_1 represents the risk difference in the binomial model. To convert the risk difference to the relative risk, Eq. (12) is expressed as $\frac{\beta_0 + \beta_1}{\beta_0}$ from Taylor series expansion at $\sum \gamma_i (Z_i - \bar{Z}_i) \approx 0$. Therefore, we can approximate the relative risk as $1 + \frac{RD}{\beta_0}$.

3.3 Simulation result

3.3.1 Log-binomial model

The odds ratio obtained by logistic regression underestimated relative risk at 0.7. In contrast, 1.5 and 3.0 were overestimates. This overestimation became bigger as incidence increased. When we set relative risk at 1.5, the odds ratio became even more exaggerated. The logistic regression standard errors were smaller than those from log-binomial regression. For the coverage probability of a 95% confidence interval, higher incidence resulted in a lower coverage rate.

Relative risks obtained from log-binomial regression were almost the same as the true relative risks. However, log-binomial regression presented convergence problems. The method could converge up to an incidence of 30%, with a relative risk of 0.7 to 1.5. A relative risk at

3.0 could be simulated up to an incidence of 20%. Standard errors from log-binomial regression were greater than those from logistic regression. For the coverage probability of a 95% confidence interval, most estimates were above 90%.

Poisson and modified Poisson regression analyses also produced almost identical true relative risks. When comparing standard errors, the modified Poisson regression yielded smaller standard errors than Poisson regression. For the coverage probability of 95% confidence intervals, Poisson and modified Poisson regression had good coverage rates.



Table1.Simulation Results: Log-binomial Model, n=1000

True RR	β_0	Logistic					Log-Binomial				Poisson				Modified Poisson		
		OR	SD	MSE	CR	RR	SD	MSE	CR	RR	SD	MSE	CR	RR	SD	MSE	CR
0.7	0.05	0.686	0.315	0.310	0.953	0.714	0.293	0.446	0.761	0.714	0.293	0.301	0.965	0.714	0.293	0.291	0.960
	0.10	0.657	0.225	0.226	0.949	0.694	0.190	0.411	0.900	0.694	0.190	0.212	0.972	0.694	0.190	0.197	0.966
	0.15	0.649	0.198	0.190	0.925	0.701	0.159	0.388	0.932	0.701	0.159	0.170	0.966	0.701	0.159	0.154	0.950
	0.20	0.633	0.170	0.170	0.908	0.705	0.134	0.383	0.959	0.698	0.135	0.148	0.966	0.698	0.135	0.130	0.943
	0.25	0.607	0.161	0.158	0.845	0.698	0.113	0.376	0.979	0.698	0.113	0.134	0.977	0.698	0.113	0.114	0.940
	0.30	0.584	0.147	0.148	0.796	0.702	0.094	0.366	0.994	0.703	0.095	0.122	0.990	0.703	0.095	0.100	0.960
	0.35	0.564	0.148	0.145	0.658	NA	NA	NA	NA	0.700	0.086	0.114	0.990	0.700	0.087	0.089	0.946
	0.40	0.532	0.145	0.141	0.513	NA	NA	NA	NA	0.699	0.080	0.104	0.989	0.699	0.080	0.077	0.946
1.5	0.05	1.568	0.312	0.316	0.962	1.526	0.294	0.515	0.835	1.528	0.293	0.308	0.962	1.484	0.293	0.290	0.957
	0.10	1.562	0.230	0.228	0.947	1.508	0.201	0.457	0.913	1.508	0.201	0.216	0.967	1.504	0.201	0.197	0.957
	0.15	1.627	0.198	0.189	0.930	1.506	0.159	0.439	0.967	1.506	0.159	0.176	0.971	1.507	0.159	0.158	0.956
	0.20	1.692	0.175	0.170	0.887	1.500	0.135	0.427	0.983	1.501	0.135	0.151	0.977	1.517	0.135	0.130	0.955
	0.25	1.759	0.156	0.158	0.833	1.498	0.113	0.419	0.996	1.500	0.113	0.134	0.984	1.501	0.113	0.114	0.963
	0.30	1.840	0.151	0.148	0.730	1.504	0.100	0.420	0.997	1.504	0.100	0.122	0.93	1.502	0.100	0.100	0.954
	0.35	1.976	0.146	0.145	0.512	NA	NA	NA	NA	1.502	0.092	0.114	0.994	1.502	0.092	0.089	0.947
	0.40	2.123	0.140	0.141	0.305	NA	NA	NA	NA	1.502	0.082	0.105	0.989	1.501	0.082	0.077	0.949
3.0	0.05	3.316	0.339	0.339	0.957	3.114	0.321	1.179	0.995	3.111	0.321	0.333	0.967	3.111	0.321	0.325	0.963
	0.10	3.463	0.254	0.243	0.920	3.068	0.222	1.142	1.000	3.065	0.222	0.232	0.963	3.065	0.222	0.221	0.949
	0.15	3.811	0.208	0.202	0.791	3.010	0.170	1.075	1.000	3.010	0.170	0.187	0.975	3.010	0.170	0.176	0.962
	0.20	4.238	0.185	0.179	0.514	3.016	0.146	1.113	1.000	3.016	0.146	0.161	0.963	3.016	0.146	0.148	0.949
	0.25	4.860	0.166	0.167	0.150	NA	NA	NA	NA	3.004	0.126	0.144	0.981	3.004	0.126	0.126	0.952
	0.30	5.812	0.156	0.158	0.011	NA	NA	NA	NA	3.013	0.111	0.130	0.983	3.013	0.111	0.114	0.954
	0.35	7.257	0.160	0.158	0.010	NA	NA	NA	NA	3.007	0.103	0.122	0.980	3.007	0.103	0.100	0.948
	0.40	9.660	0.166	0.161	0.010	NA	NA	NA	NA	3.016	0.091	0.114	0.986	3.016	0.091	0.094	0.955

Estimate is the mean of the parameter estimates based on 1,000 replicates);SD is the empirical standard deviation of the parameter estimate; MSE is the mean value of the estimated standard errors; CR is the coverage probability; NA means failed to converge.

3.3.2 Binomial model

The odds ratio obtained by logistic regression was slightly greater than the true relative risk at 0.7, within an incidence rate of 25%, and true relative risk at 1.5 with an incidence rate from 20 to 40%. Otherwise, logistic regression produced smaller odds ratios. The overestimation became more critical at a true relative risk of 3.0, with a high incidence rate. For the coverage rate, the higher the incidence rate, the lower the coverage rate it produced because the estimate ratio was biased. There was no difference in the standard error among the scenarios considered.

Relative risk from log-binomial regression was smaller than the true relative risk at 1.5 and 3.0. In contrast, log-binomial regression gave a higher relative risk than the true relative risk of 0.7. It had convergence problems, as discussed for the log-binomial model. It only failed to converge at a true relative risk of 3.0. The standard errors for the coverage rate did not differ.

Poisson and modified Poisson regression analyses overestimated relative risk at a true relative risk of 0.7; otherwise, they underestimated relative risk. Using sandwich variance estimates to compute standard errors, the modified Poisson regression produced smaller standard errors than the ordinary Poisson regression. For the coverage rate, both methods provided lower coverage rates than the other methods. These results indicated that regular Poisson regression produced a higher coverage rate than modified Poisson regression.

Table2.Simulation Results: Binomial Model, n=1000

True RR	Incidence	Logistic				Log-Binomial				Poisson				Modified Poisson			
		OR	SD	MSE	CR	RR	SD	MSE	CR	RR	SD	MSE	CR	RR	SD	MSE	CR
0.7	0.05	0.898	0.183	0.181	0.733	0.923	0.146	0.167	0.398	0.921	0.146	0.161	0.621	0.921	0.146	0.146	0.535
	0.10	0.841	0.166	0.164	0.801	0.866	0.124	0.189	0.555	0.864	0.125	0.144	0.719	0.864	0.125	0.126	0.608
	0.15	0.784	0.153	0.154	0.895	0.838	0.113	0.209	0.785	0.836	0.112	0.131	0.766	0.836	0.112	0.112	0.649
	0.20	0.738	0.148	0.144	0.927	0.817	0.104	0.225	0.786	0.815	0.104	0.121	0.800	0.815	0.104	0.100	0.667
	0.25	0.710	0.144	0.141	0.945	0.802	0.090	0.238	0.864	0.800	0.089	0.113	0.840	0.800	0.089	0.091	0.698
	0.30	0.679	0.136	0.137	0.951	0.788	0.082	0.250	0.918	0.785	0.081	0.106	0.868	0.785	0.081	0.077	0.718
	0.35	0.638	0.134	0.134	0.899	0.781	0.076	0.256	0.947	0.779	0.075	0.100	0.891	0.779	0.075	0.075	0.717
	0.40	0.609	0.137	0.134	0.821	0.770	0.068	0.270	0.978	0.767	0.068	0.095	0.920	0.767	0.068	0.069	0.739
1.5	0.05	1.192	0.177	0.176	0.723	1.158	0.139	0.202	0.536	1.160	0.139	0.156	0.629	1.160	0.139	0.140	0.558
	0.10	1.340	0.159	0.158	0.878	1.221	0.118	0.232	0.692	1.225	0.118	0.135	0.691	1.225	0.118	0.117	0.588
	0.15	1.436	0.147	0.144	0.939	1.276	0.103	0.264	0.845	1.280	0.103	0.122	0.779	1.280	0.103	0.101	0.639
	0.20	1.543	0.143	0.141	0.942	1.312	0.093	0.288	0.925	1.315	0.092	0.111	0.817	1.315	0.092	0.088	0.668
	0.25	1.685	0.141	0.137	0.849	1.336	0.079	0.300	0.977	1.341	0.078	0.103	0.873	1.341	0.078	0.078	0.673
	0.30	1.858	0.133	0.134	0.647	1.352	0.070	0.309	0.992	1.355	0.069	0.096	0.890	1.355	0.069	0.070	0.694
	0.35	2.070	0.135	0.134	0.347	1.368	0.064	0.319	0.995	1.371	0.063	0.091	0.917	1.371	0.063	0.062	0.683
	0.40	2.422	0.140	0.137	0.060	1.379	0.056	0.327	1.000	1.382	0.055	0.086	0.936	1.382	0.055	0.055	0.674
3.0	0.05	1.840	0.172	0.167	0.170	1.601	0.133	0.489	0.769	1.611	0.131	0.149	0.018	1.611	0.131	0.133	0.009
	0.10	2.643	0.155	0.151	0.853	1.904	0.112	0.183	0.994	1.913	0.111	0.127	0.050	1.913	0.111	0.109	0.027
	0.15	3.698	0.142	0.144	0.709	NA	NA	NA	NA	2.125	0.094	0.114	0.118	2.125	0.094	0.092	0.055
	0.20	5.663	0.148	0.144	0.006	NA	NA	NA	NA	2.259	0.081	0.103	0.173	2.259	0.081	0.080	0.083
	0.25	12.692	0.168	0.167	0.000	NA	NA	NA	NA	2.358	0.070	0.095	0.246	2.358	0.070	0.069	0.085
	0.30	161.90	0.482	0.472	0.000	NA	NA	NA	NA	2.422	0.060	0.089	0.294	2.422	0.060	0.060	0.093
	0.35	960000	0.110	703.83	1.000	NA	NA	NA	NA	2.496	0.057	0.084	0.383	2.496	0.057	0.054	0.125
	0.40	783000	0.109	704.04	1.000	NA	NA	NA	NA	2.539	0.051	0.080	0.431	2.539	0.051	0.050	0.121

Estimate is the mean of the parameter estimates based on 1,000 replicates); SD is the empirical standard deviation of the parameter estimate; MSE is the mean value of the estimated standard errors; CR is the coverage probability; NA means failed to converge.

IV. Illustrative data

We considered the data from a cohort. The data was collected over the period of time from 2002–2010 by the National Health Insurance Service (NHIS 2014). The total number of enrolled patients was 1,018,682 during the baseline period of 2002–2003 (NHIS 2014). We were interested in studying the relationship between obesity and diabetes. We identified diabetic patients who had diabetes after 2004, resulting in 105,091 diabetic patients. Diabetes incidence for that period was 10.32%. This data set included gender, disease, status of death, and the body mass index (BMI) of the patients. When analyzing the data, we ignored the patients who did not have BMI information. The final data represented 451,865 patients who had complete data for gender, age group, diabetes status, status of death, and BMI. Because there was no obesity status variable in the data, we defined BMI scores of under 23 as normal. Obesity status was the main independent variable, and the others were covariates. The summary of data is provided in Table 3.

In this study, we built a log-binomial model and compared it to the three other models (logistic, Poisson, and modified Poisson regression), using the same predictors and outcomes. The regression analysis was based on the final data ($n = 451,865$), with a diabetes incidence rate fixed at 10.32%.

Table 3. A summary of illustrative data

	Total		Diabetes		Normal	
	N	%	N	%	N	%
Obesity						
Yes	242,599	53.63	41,615	17.15	200,984	82.85
No	209,775	46.37	21,837	10.41	187,938	89.59
Death						
Yes	97	0.02	15	15.46	82	84.54
No	452,277	99.98	63,437	14.03	388,840	85.97
Age						
0-9	889	0.20	23	2.59	866	97.41
10-19	52,261	11.55	1,828	3.50	50,433	96.50
20-29	88,731	19.61	4,641	5.23	84,090	94.77
30-39	118,035	26.09	11,483	9.73	106,552	90.27
40-49	94,987	21.00	17,426	18.35	77,561	81.65
50-59	58,674	12.97	15,851	27.02	42,823	72.98
60-69	31,135	6.88	9,979	32.05	21,156	67.95
≥70	7,662	1.69	2,221	28.99	5,441	71.01
Sex						
Male	228,649	50.54	30,964	13.54	197,685	86.46
Female	223,725	49.46	32,488	14.52	191,237	85.48

Application of the logistic regression procedure resulted in an estimated odds ratio of 1.469 (95 percent CI: 1.443–1.497); this value differed significantly from the results obtained using log-binomial regression given the 1.356 estimated relative risk (95% CI: 1.3367–1.3773). Using Poisson regression analysis resulted in an estimated relative risk of 1.364 (95% CI: 1.3418–1.3872); again, this risk differed from the estimated relative risk from log-binomial regression, with a slightly higher relative risk. The estimated relative risk from modified Poisson regression was the same as that from Poisson regression analysis, but it gave smaller standard errors than Poisson regression (95% CI: 1.3439–1.3851). A summary of the results is provided in

Table 4.

Table4. Result of the estimated odds ratio and relative risk by different regression model.

Method	OR or RR	SE	95% CI
Logistic regression	1.470	0.004	1.443, 1.497
Log-binomial regression	1.356	0.010	1.336, 1.377
Poisson regression	1.364	0.011	1.341, 1.387
Modified Poisson regression	1.364	0.010	1.343, 1.385

OR is the odds ratio; RR is relative risk SE is standard error; 95% CI is 95 percent of confidence interval.



V. Conclusion and Discussion

In this paper, we proposed different methods to estimate relative risk in a binary response variable. The odds ratio could be directly obtained by logistic regression. However, the odds ratio should not be replaced with the relative risk in cohort studies under some conditions. Converting odds ratios to relative risks could produce overestimates or underestimates under some conditions, particularly with the incidence increasing. The overestimation or underestimation could exaggerate the treatment effects in a study. Therefore, proper data analysis methods should be used.

Through the results of the simulation, the estimated relative risks in the log-binomial model provided good performance when Poisson and modified Poisson regression were applied. Modified Poisson regression analysis produced lower MSEs when using sandwich variance estimates. Log-binomial regression gave results similar to those from Poisson and modified Poisson regression analyses, except for convergence problems. The reasons of the convergence problems were the failure to find the maximum likelihood estimate (MLE). The log-binomial model was placed on the boundary of the parameter space, and the log-likelihood function was maximized on the boundary of the parameter space (Williamson, Eliasziw, and Fick 2013), and it might also happen with many covariates, especially continuous covariates. The convergence problem could be avoided using the COPY method in SAS (Lumley, Kronmal, and Ma 2006) or a different method such as modified Poisson regression, which outperformed the other

methods in terms of estimating relative risk, MLE, and convergence.

We applied these data analysis methods to data from a diabetes study. The odds ratio obtained from logistic regression and the relative risk were different due to high incidence. Thus, the odds ratio was not a good estimate of relative risk. Application of binomial regression had the smallest adjusted relative risk compared with the other regression analyses. As we expected, using modified Poisson regression analysis yielded a smaller standard error than Poisson regression.

Different data analysis methods provided different relative risk estimates. Moreover, with high incidence and a typical outcome, the odds ratio obtained from logistic regression provided large differences. In this case, alternative methods should be considered, as logistic regression led to an exaggerated or underestimated risk association or treatment effect. Thus, determining the method to estimate adjusted relative risk is important. There are limitations in the binomial model. The estimate ratio was generally biased and required a correction method. It was difficult to choose the best correction method.

Bibliography

Hosseinian, S., 2009. Robust inference for generalized linear models: binary and poisson regression. PhD. Thesis, Ecole Polytechnique Federal de Lausanne.

Knol, M. J., Le Cessie, S., Algra, A., Vandenbroucke, J. P., & Groenwold, R. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *CMAJ*, 184(8), 895-899. doi: 10.1503/cmaj.101715

Lumley T, Kronmal RA, Ma S. Relative risk regression in medical research: models, contrasts, estimators and algorithms. Seattle, Wash: University of Washington; 2006. UW Biostatistics Working Paper Series, paper 293. Available at: <http://www.bepress.com/uwbiostat/paper293>. Accessed July 19, 2006.

Spiegelman, D., & Hertzmark, E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol*, 162(3), 199-200. doi: 10.1093/aje/kwi188

Williamson, T., Eliasziw, M., & Fick, G. H. (2013). Log-binomial models: exploring failed convergence. *Emerg Themes Epidemiol*, 10(1), 14. doi: 10.1186/1742-7622-10-14

Winkelmann, R. (2003). *Econometric Analysis of Count Data*: Springer.

Zhang, J., & Yu, K. F. (1998). What's the relative risk? A method of correcting

the odds ratio in cohort studies of common outcomes. JAMA, 280(19), 1690–1691.

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol, 159(7), 702–706.

Zou, G. Y., & Donner, A. (2013). Extension of the modified Poisson regression model to prospective studies with correlated binary data. Stat Methods Med Res, 22(6), 661–670. doi: 10.1177/0962280211427759



국문요약

이분형 반응변수에서 상대위험도

추정방법에 대한 비교

오즈비 및 상대위험도는 보건분야 및 임상시험에서 많이 쓰이는 지표이다. 환자-대조군 연구에서 오즈비는 로지스틱회귀분석을 통해 얻어질수 있다. 하지만 코호트 연구에서는 오즈비를 상대위험도로 대체되어 사용되어지게 되면 연구결과를 과대평가 또는 과소평가로 이어질수도 있다. 본 연구에서는 이분형 반응변수에서 상대위험도를 추정하는 몇가지 방법에 대하여 알아보하고자 하였다. 오즈비는 로지스틱회귀분석에서 추정된다. 발생률이 10%가 넘을때에는 오즈비가 상대위험도로 대체되어 사용되어서는 안된다. 이때에는 상대위험도를 추정하는 다른 방법들을 사용해야 한다. 로그 바이노미얼 회귀분석은 로지스틱을 대체하여 상대위험도를 추정하는 방법이다. 하지만 로그-바이노미얼 회귀분석방법은 수렴을 하지 못하는 단점을 가지고 있다. 이러한 단점을 보완하는 로버스트 포아송 회귀를 통한 상대위험도 추정방법이 있다. 로버스트 포아송 회귀분석에서 추정도 상대위험도의 표준오차는 일반적인 포아송 회귀분석에서 추정된 상대위험도의 표준오차에 비하여 작다. 모의시험결과 발생률에 따른 오즈비 및 상대위험도를 추정하였는데 로버스트 포아송회귀에서 대체적으로 좋은 결과가 나왔다. 하지만 모의시험에서 바이노미얼 모델에서 추정된 위험차를 상대위험도로 변환하여 생각하였는데 한계점이 나타났다. 바이노미얼 모델에서 위험차를 상대위험도로 바꾸었을 때 추정하는 방법에 대하여 추후 연구가 필요하다. 실제 예제 데이터를 이용하였을 때도 로버스트 포아송회

귀 분석에서 가장 좋은 결과를 보였다.



핵심 되는 말 : 오즈비, 상대위험도, 로지스틱 회귀분석, 로그-바이노미얼 회귀분석, 포아송 회귀분석, 로버스트 포아송 회귀분석, 로그-바이노미얼 모델, 상대위험도 추정방법