Log-binomial Regression Models for HTA Submissions

Chi-Shian Dai[1], Richard Baumgartner[2], Shahrul Mt-lsa[3], William Malbecq[4]

[1] University Wisconsin-Madison, Madison, WI, USA

[2] Biostatistics and Research Decision Sciences, Merck & Co., Kenilworth, NJ, USA

[3] Biostatistics and Research Decision Sciences, MSD, Zurich, Switzerland

[4] Biostatistics and Research Decision Sciences, MSD, Brussels, Belgium

08-26-2022

Abstract

There is a high demand for binary data modeling in the health technology assessment (HTA) applications. Specifically, BARDS-HTA Statistics summarizes binary data through risk ratios in compliance with guidance documents from the HTA agencies. In this case, the log-binomial regression is more suitable than logistic regression for these analyses. However, it is well-known that the log-binomial regression has several limitations, including a high non-convergence rate, specifically in the presence of multiple covariates and due to extreme event counts. The non-convergence issue can be addressed by adding constraints to the log-binomial regression. In this project, we compared the constrained log-binomial regression with other binary regression methods in following settings: (1) with multiple covariates, (2) analysis of dependent variables with low or high proportions, and (3) longitudinal data setting. We discuss the potential application of the constrained log-binomial regression in cross-sectional analyses, and then outline the challenges in analyzing longitudinal data to inform future research.

Keywords— log-binomial models, risk ratio, generalize estimating equation, longitudinal study, cross-sectional study

1 Introduction

There is a high demand for binary data modeling in health technology assessment (HTA) applications. Traditionally, statisticians summarize binary response data through logistic regression, which has been well developed for over a hundred years. However, the logistic regression provides an odds ratio that is hard to interpret for scientists. In this case, the log-binomial regression is more suitable than logistic regression for these analyses since the coefficient of the log-binomial regression indicates the risk ratio. However, it is well-known that the log-binomial regression has several limitations, including a high non-convergence rate (Williamson et al. 2013), specifically in the presence of multiple covariates and due to extreme event counts.

Zou (2004) approach the risk ratio via a Poisson regression which brings several benefits, including that there is no convergence issue and it is easy to implement. However, the estimated probability from a Poisson regression may be greater than one, which does not make sense for scientists. Furthermore, since Poisson regression is a wrong model for binary data, it cannot reach Cramér–Rao lower bound. And hence, Poisson regression is not optimal. We favor log-binomial regression over Poisson regression since it can reach Cramér–Rao lower bound and gives a probability estimate. But, the convergence issue of the log-binomial model needs to be solved.

Luo et al. (2014) proposed a log-binomial model with constraints(LBC) that can handle the convergence issue of the log-binomial model. Schwendinger et al. (2021) shows that conic programming has the most stable performance by comparing several optimization solvers for LBC. However, little did people know how the performance of LBC compares to the Poisson regression. In the first part of this paper, we conduct a simulation study to evaluate the performance of these two methods in the following settings: (1) with multiple covariates, (2) analysis of dependent variables with low or high proportions.

How to implement log-binomial models for longitudinal data is still an open problem. In the second part of this paper, we consider a generalized estimating equations type of log-binomial models for longitudinal data(LBC-GEE). Same as the cross-sectional study, the packages **gee** and **geepack** in **R** cannot converge while applying log-binomial models. We proposed a reclusive algorithm for LBC-GEE, which solves the convergence issue. However, we face more challenges, such as computationally heavy and consistent estimates for standard deviation.

We organize this paper as follows. Section 2 introduces the log-binomial models with constraints (LBC) and compares its performance with Poisson regression in a cross-sectional study. Section 3 presents the generalized estimating equation type of log-binomial models(LBC-GEE) for the longitudinal data. In section 4, we apply the methodologies LBC and LBC-GEE to simulated clinical trial data. Section 5 contains discussion and summary. Section 6 gives an introduction to the supplementary documents.

2 Log-binomial Models with Constraints for Cross-sectional Study

Let y_i be the response, $\boldsymbol{x}_i \in \mathbf{R}^p$ is the covariate variable. The log-binomial model is a generalize linear regression model with a binary family and the log link function. And hence, the model is

$$\log P(Y_i = 1|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

From this we can calculate the risk ratio,

$$P(Y_i = 1|\boldsymbol{x}_i)/P(Y_i = 1|\boldsymbol{x}_i') = \exp\{(\boldsymbol{x}_i - \boldsymbol{x}_i')^\top \boldsymbol{\beta}\}.$$

And hence, the coefficient $\boldsymbol{\beta}$ indicates the risk ratio. Maximal likelihood estimate are used to estimate the coefficient $\boldsymbol{\beta}$.

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \tag{1}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\boldsymbol{x}_i^\top \beta) + (1 - y_i)\log\{1 - \exp(\boldsymbol{x}_i^\top \beta)\}, \tag{2}$$

where ℓ is the log-likelihood function. From (2), we can see that the reason of the convergence issue. In order to solve (1), a typical way is to apply algorithms for convex optimization,, for example: gradient decent or Newton's method, which give a sequence $(\boldsymbol{\beta}_1, \dots \boldsymbol{\beta}_k \dots)$ and eventually this sequence can approach the optimizer $\widehat{\boldsymbol{\beta}}$. However, the element in this sequence may make $\exp(\boldsymbol{x}_i^\top \beta_k)$ greater than 1. And hence, the log term in (2) would take a negative value which is not feasible. As a result, it may not converge.

Luo et al. (2014) proposed a log-binomial models with constraints(LBC) which can solve the convergence issue of log-binomial models.

$$\begin{aligned} \widehat{\gamma} &= \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \\ &\text{subject to} \quad \boldsymbol{x}_i^\top \boldsymbol{\beta} < 0 \quad \forall i = 1, \dots, n. \end{aligned} \tag{3}$$

The constraints (3) can make sure the exponential term in (2) is less than 1.

- In this paper, we implement LBC via conic programming proposed by Schwendinger et al. (2021).

- The constraints in (3) are implemented by

$$\boldsymbol{x}_i^\top \boldsymbol{\beta} \le -0.1^7$$

since the constraints need to be closed sets.

2.1 Adjusted Confidence Intervals

Since there are constraints in (3), we cannot use fisher information matrix to estimate standard deviation. de Andrade & Andrade (2018) proposes an adjusted confidence intervals for MLE with constraints.

1. Calculate $\widehat{\boldsymbol{\beta}}$, and the Fisher information matrix $I(\boldsymbol{\beta})$.
2. Let A be the matrix that collects the rows of the designed matrix X satisfying $\boldsymbol{x}_i^\top \widehat{\beta} = 0$.
3. We have the following asymptotic theory de Andrade & Andrade (2018).

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = I^{-1} - I^{-1}A'(AI^{-1}A')^{-1}AI^{-1}. \tag{4}$$

2.2 Simulation

In this subsection, we conduct a simulation study to show that two things. First, the constraints can solve the convergence issue of log-binomial models. Second, to compare the performance between Poisson regression and Log-binomial model.

We generate covariates $\boldsymbol{X} = (X_1, \dots, X_p)$ from i.i.d. uniform distribution from (0,1). And, a treatment covariate from a randomized controlled trial with a probability of 0.5. And, we generate a response following a log-binomial models.(There is no model misspecification)

$$\log P(Y = 1|\boldsymbol{X}, Trt) = \log(3)Trt + c_0 - \sum_{j=1}^{p} 0.5^p X_j.$$

We implement three methods.

- Log-binomial(LB): **glm** function with binomial family and *log* link in **R**.

- Log-binomial with constraints (LBC): Conic programming. **ROI** package in **R**Schwendinger et al. (2021).
- Poisson Regression: **glm** function with Poisson family and *log* link in **R**.

Figure (1) shows the convergence rate of three methodologies. LB may not converge, especially when the sample size n is small and the dimension p is large. On the other hand, LBC and Poisson regression have similar convergence rates. And hence, it shows that LBC solves the convergence issue of LB. Figure (2) shows the coverage rate of 95% confidence interval for the log-risk ratio. If the line is closer to 0.95, the method is better. LBC is closer to 0.5 than Poisson regression when the sample size n is small; dimension p is large, or the success rate is low.

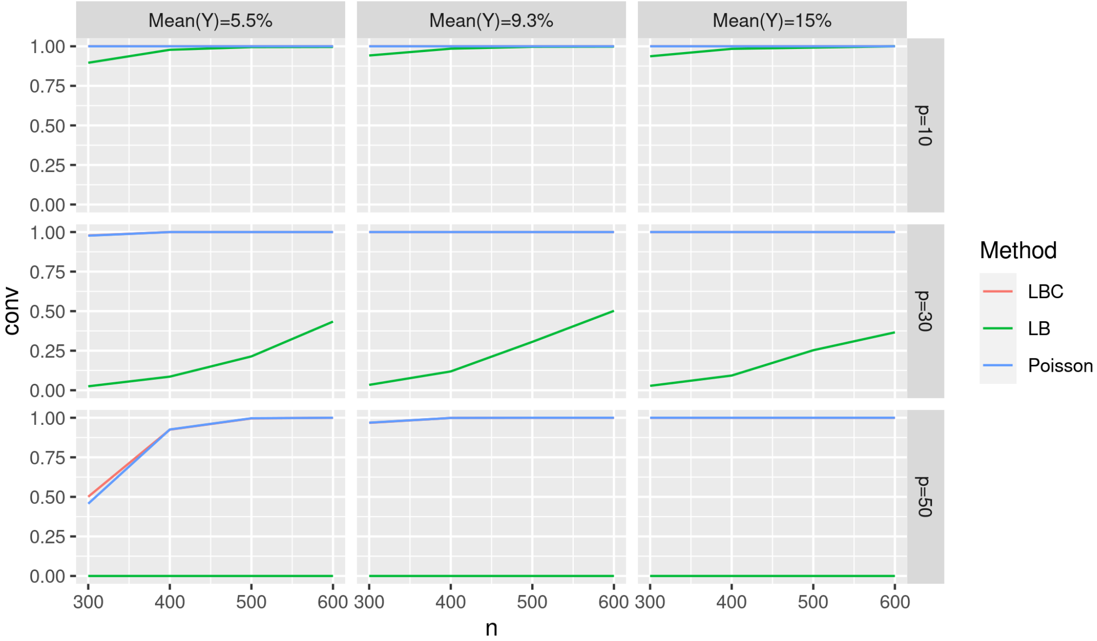


Figure 1: Convergence rate. "conv" denotes the convergence rate. n is sample size. p is the dimension of covariates. "Mean(Y)" is the average of the response. The number of the replications is 1000.

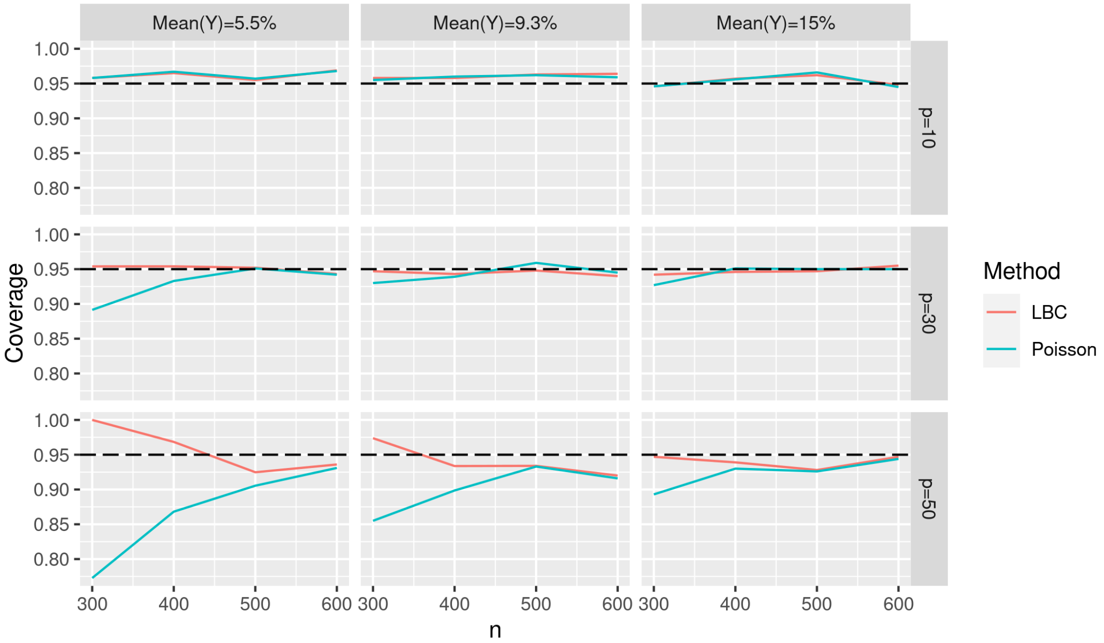


Figure 2: Coverage rate of 95% confidence interval. ”Coverage” denotes the coverage rate. n is sample size. p is the dimension of covariates. ”Mean(Y)” is the average of the response. The number of the replications is 1000.

3 Log-binomial Models for Longitudinal Study

Generalize estimating equation(GEE) is a well-known approach to longitudinal study Wang (2014). To implement GEE types of log-binomial models, the existing packages **gee** and **geepack** in **R** fail to converge. On the other hand, the existing package **gee** and **geepack** only consider the constant correlation structure which are not feasible for binary response data Rao Chaganty & Joe (2004). This section proposes a recursive algorithm for GEE type of LBC that solves the above two existing packages' problems.

Let $\boldsymbol{y}_i = (y_{i1}, \dots, y_{im})$ be the repeated measurement for patient i. $\boldsymbol{x}_{ij} \in \mathcal{R}$ be the covariate for patient i at time point j. GEE estimate is the solution of the following equations.

$$\sum_{i}^{n} \nabla \boldsymbol{\mu}_i(\boldsymbol{\beta})^\top V_i^{-1}(\boldsymbol{\beta})\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = 0, \qquad (5)$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \dots, \mu_{im}(\boldsymbol{\beta})$ is the models for $\boldsymbol{y}_i$, $\mu_{ij} = P(y_{ij} = 1|\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) = \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})$, and V_i is the variance of $\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})$. We would like to borrow the idea from cross-sectional study, add constraints to the equations (5). The simple approach is adding constraints to the integrals of the equations (5). However, since the variance V is depend on $\boldsymbol{\beta}$ the integrals can be complicated which cannot be derived. But, if we have a constant estimate $\widehat{V}_i$ for $V(\boldsymbol{\beta}^*)$, the integral has the following form

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i^n \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\}^\top \widehat{V}_i^{-1} \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\},$$

which is a weighted least square. And hence, we proposed the following algorithm

Step 1: Ignore the longitudinal structure. Get $\widehat{\boldsymbol{\beta}}_0$ by the standard LBC.

Step 2: Use $\widehat{\boldsymbol{\beta}}_k$ to estimate $\widehat{V}_i$.

Step 3: Get the new estimate $\widehat{\boldsymbol{\beta}}_{k+1}$ by solving the following constrained optimization problem.

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = &\arg\min_{\boldsymbol{\beta}} \sum_i^n \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\}^\top \widehat{V}_i^{-1} \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\}, \\ \text{subject to} \quad & \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} < 0 \qquad \forall\ i, j. \end{aligned}$$

Step 4: Repeat step 2 and step 3 until $\widehat{\boldsymbol{\beta}}_k$ reaches the stopping rule.

3.1 How to estimate V_i?

From Wang (2014), $V_i = A_i^{1/2} R_i A_i^{1/2}$, where $R_i = \mathrm{Cor}(\boldsymbol{y}_i, \boldsymbol{y}_i)$ is the correlation structure, $A_i = diag(v_{i1}, \ldots, v_{im})$ is the variance of the models. In the binary data, $v_{ij} = \widehat{p}_{ij}(1 - \widehat{p}_{ij})$ and $\widehat{p}_{ij} = \exp(\boldsymbol{x}_{ij}^\top \widehat{\boldsymbol{\beta}})$. And hence the only thing unknown is the correlation structure. The traditional ways is using a constant correlation structure, ex: "unstructured" and "exchangeable" (see Wang (2014)). However, the constant correlation structure is not appropriate for binary data since the domain of the correlation is not (-1,1) Rao Chaganty & Joe (2004). The following shows the domain of $\mathrm{Cor}(y_{ij}, y_{ik})$

$$-\sqrt{\frac{P_{ij}P_{ik}}{(1-P_{ij})(1-P_{ik})}} \leq \mathrm{Cor}(y_{ij}, y_{ik}) \leq \frac{\min\{P_{ij}, P_{ik}\} - P_{ij}P_{ik}}{\sqrt{P_{ij}P_{ik}(1-P_{ij})(1-P_{ik})}}, \tag{6}$$

where $P_{ij} = P(Y_{ij} = 1)$, $P_{ik} = P(Y_{ik} = 1)$. And hence, if it was a constant correlation structure, the constant can be outside the feasible domain. Inspired by Rao Chaganty & Joe (2004), it suffices to estimate the join distribution $P(Y_{ij}Y_{ik} = 1)$ since

$$\mathrm{Cor}(y_{ij}, y_{ik}) = \frac{P(Y_{ij}Y_{ik} = 1) - P_{ij}P_{ik}}{\sqrt{P_{ij}P_{ik}(1 - P_{ij})(1 - P_{ik})}},$$

where the marginal probability P_{ij} and P_{ik} are given by the estimate $\widehat{\boldsymbol{\beta}}_k$ from the previous iteration. The following algorithm provides a possible way to estimate the joint distribution.

Step 1: Assume that the binary responses $(Y_{i1}, \ldots Y_{im})$ comes from dichotomizing a multivariate normal distribution with an exchangeable correlation.

Step 2: We have the estimated marginal probability $\widehat{p}_{ij}$ derived by the estimate $\widehat{\boldsymbol{\beta}}_k$ from the previous iteration.

Step 3: Find the normal quantiles q_{ij}.

$$P(Z < q_{ij}) = \widehat{p}_{ij}$$

Step 4: Joint probability would be

$$P(Y_{ij}Y_{ik} = 1 | \widehat{p}_{ij}, \widehat{p}_{ik}) = P(Z_1 < q_{ij}, Z_2 < q_{ik} | \rho),$$

where (Z_1, Z_2) is a bivariate normal distribution with mean 0, variance 1, covariance ρ.

Step 5: Since the joint probability is a function of ρ, we can find the estimate via MLE.

The different assumption in Step 1 gives a different likelihood in Step 5. For example, it can be an AR(1) correlation for multivariate normal or another distribution like a multivariate gamma. This paper only considers the simplest case, which is presented in Step 1.

3.2 How to estimate variance?

From sandwich method provides a consistent estimate of variance for GEE.

- Sandwich Covariance Estimator:

$$
\begin{aligned}
V =& \left\{\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right\}^{-1} M \left\{\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right\}^{-1}, \\
D_i =& \nabla \boldsymbol{\mu}_i, \\
M =& \sum_{i=1}^{n} D_i^\top V_i^{-1} Cov(Y_i) V_i^{-1} D_i, \\
Cov(Y_i) =& (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i)^\top.
\end{aligned}
$$

- Adjusted Sandwich Covariance Estimator: Same as the cross-sectional study, we need to consider an adjustment on the covariance matrix since we add the constraints.

  1. Calculate $\widehat{\beta}$, and sandwich covariance estimator V.
  2. Let A be the matrix formed with the $q \leq n$ rows of $\boldsymbol{X}$ which $\boldsymbol{x}_i^\top \widehat{\beta} = 0$.
  3. Adjusted Sandwich Covariance Estimator is

$$
V_a = V^{-1} - V^{-1} A' (A V^{-1} A')^{-1} A V^{-1}.
$$

3.3 Stopping Rule

The stopping rule is also crucial for this algorithm. In theory, $\widehat{\boldsymbol{\beta}}_k$ is consistent for every k since $\widehat{\boldsymbol{\beta}}_K$ is consistent no matter the correlation structure is correct or not. So, a naive way is to let k equal a fixed number. However, intuitively we would like to keep iteration until the estimate $\widehat{\boldsymbol{\beta}}_k$ converges. However, we realize that there is around 3% in the simulation diverges. While we look at the estimate sequence $\widehat{\boldsymbol{\beta}}_k$, $\widehat{\boldsymbol{\beta}}_k$ are stable in the early steps but suddenly have a huge jump which causes the divergence. Since $\widehat{\boldsymbol{\beta}}_k$ is consistent for every k, we would stop the iteration if the difference between $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\beta}}_{k+1}$ is huge. Please see the document "LBC-GEE.html" for further detail.

- There is around 3% in the simulation that it diverges. $\widehat{\boldsymbol{\beta}}_k$ are stable in the early steps. To deal with it, we consider an "Early-stop".
- Early-stop: Stop if the difference between $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\beta}}_{k+1}$ is huge. Then, return $\widehat{\boldsymbol{\beta}}_k$.
- No-stop: It reaches the maximal iteration number (20).

3.4 Simulation for Longitudinal Data

In this simulation, we consider two methodologies. Our proposed algorithm is LBC-GEE. And the GEE type of Poisson regression is implemented by the **geeglm** function in the package **geepack** with an exchangeable correlation. We generate the response following the log-binomial models; hence, there is no model misspecification. Here is the detail of how we generate the data.

- Covariates: $\boldsymbol{X} = (X_1, \ldots, X_p)$ *i.i.d.* $Unif(0, 1)$.
- Treatment:

$$T_k \sim Ber(0.5)$$

- Generate response Y_{ij} $j = 1, \ldots, m$

1. For $j = 1, \ldots, m$, calculate the probabilities.

$$\begin{aligned}\log P_{ij} &= \log p(Y_{ij} = 1 | T_i, \boldsymbol{X}_i) \\ &= aT_i - \log(3) - c - 2 * \sum_{j=1}^{p} 0.5^p \boldsymbol{X}_{ij},\end{aligned}$$

2. Calculate the quantiles q_{ij} of probabilities P_{ij} form the standard normal distribution.

$$p(Z \leq q_{ij}) = P_{ij}.$$

3. Generate multivariate normal distribution $(Z_1, Z_2, \ldots, Z_m)$ with mean 0, variance 1, $\mathrm{Cor}(Z_i, Z_j) = 0.6$.

4. Let $Y_{ij} = 1$ if $Z_{ij} \leq q_{ij}$, otherwise $Y_{ij} = 0$.

We dichotomize a multivariate normal distribution to create a coherence of repeated measurements.

Table 1 shows the bias and the coverage rates of 95% confidence intervals for the log-risk ratio at a single time point (visit 1 and visit 5). One can see that LBC-GEE and Poisson-GEE are both consistent estimates. And LBC-GEE and Poisson-GEE have similar performance while the success rate is low. But, if the success rate is high, LBC-GEE provides a non-valid confidence interval. So, our first challenge is how to estimate standard deviation.

The relative mean square error is a well-known way to evaluate the performance of two methods. We first calculate the mean square error with the following formula.

$$\frac{1}{T}\sum_{j=1}^{T}(\text{Estimated log risk ratio at simulation j} - \text{True log risk ratio})^2,$$

where T is the number of the replication. And, the relative mean square error(%) is

$$\frac{MSE(Poisson) - MSE(LBC)}{MSE(LBC)} \times 100.$$

If the value is negative, that means Poisson-GEE is better. Otherwise, LBC-GEE is better. Table 2 shows the relative mean square error(%). In general, LBC-GEE and Poisson-GEE have similar performances. But, in some cases, the improvement of LBC-GEE can be up to 4 %, which indicates that if we have a consistent estimate for standard deviation, it could be a smaller length of the confidence interval. And hence, it has a larger power.

	Mean(Y)=		8%		13%		23%		31%	
n			LBC	Poi	LBC	Poi	LBC	Poi	LBC	Poi
200	visit1	bias	0.04	0.04	0.02	0.02	0.02	0.02	0.01	0.02
		Coverage	0.97	0.97	0.96	0.96	0.89	0.94	0.85	0.95
	visit5	bias	0.09	0.09	0.03	0.03	0.02	0.02	-0.02	0.03
		Coverage	0.96	0.96	0.95	0.95	0.95	0.95	0.93	0.95
300	visit1	bias	0.03	0.03	0.03	0.03	0.00	0.00	-0.00	0.01
		Coverage	0.96	0.96	0.94	0.94	0.91	0.94	0.84	0.96
	visit5	bias	0.03	0.03	0.02	0.02	0.01	0.01	-0.03	0.02
		Coverage	0.96	0.96	0.95	0.95	0.95	0.95	0.93	0.95

Table 1: 95% confidence intervals, $p = 2$, $m = 5$. LBC denotes our proposed algorithm LBC-GEE. Poi denotes the GEE type of Poisson regression. The number of the replications is 800.

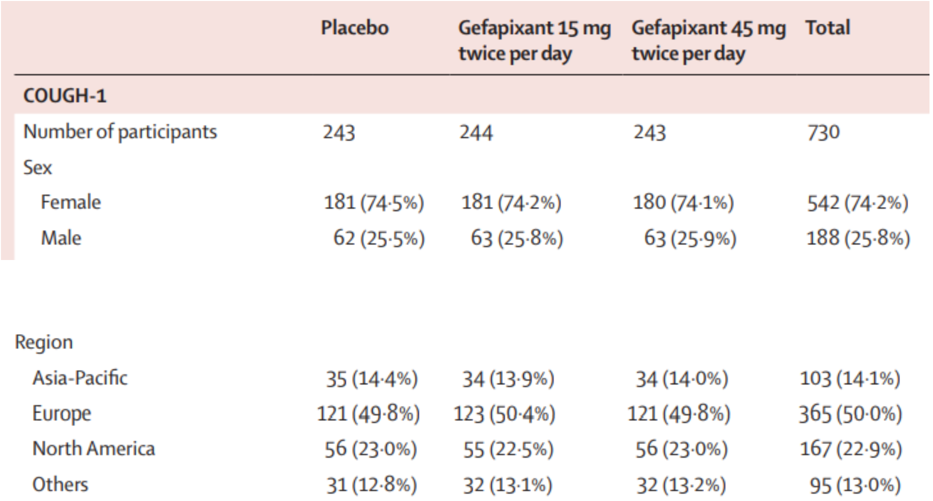
n		Mean(Y)=8%	Mean(Y)=13%	Mean(Y)=23%	Mean(Y)=31%
200	visit1	-0.73	-0.36	-0.01	-0.21
	visit5	-0.37	0.69	1.53	2.55
300	visit1	0.20	0.08	0.67	-0.91
	visit5	0.15	0.21	3.99	0.47

Table 2: Relative Mean Square Error (%). LBC denotes our proposed algorithm LBC-GEE. Poi denotes the GEE type of Poisson regression. The number of the replications is 800.

4 Simulated Clinical trial Data

In this section, we generate a simulated clinical trial data from McGarvey et al. (2022). In the previous simulation, we generate the response following the log-binomial models. But, here, we generate the response following the clinical trial data. And hence, there may have model misspecification. This clinical trial study wants to evaluate the efficacy of gefapixant for chronic cough. It is a longitudinal study with three visits and 730 patients. The response is 24-h cough frequency. And we dichotomize the response by 30% reduction or more in 24-h cough frequency. There are three treatments, placebo, 15mg, and 45mg. Three covariates ”Sex”, ”Region”, ”Baseline”.(24-h cough frequency in the first day) Here is the detail of how we generate this data set.

- Generate categorical covariates: Treatment Sex and Region with the following marginal distribution.

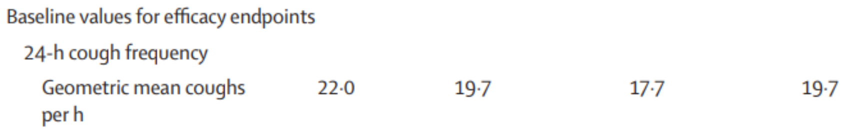


	Placebo	Gefapixant 15 mg twice per day	Gefapixant 45 mg twice per day	Total
COUGH-1				
Number of participants	243	244	243	730
Sex				
Female	181 (74·5%)	181 (74·2%)	180 (74·1%)	542 (74·2%)
Male	62 (25·5%)	63 (25·8%)	63 (25·9%)	188 (25·8%)
Region				
Asia-Pacific	35 (14·4%)	34 (13·9%)	34 (14·0%)	103 (14·1%)
Europe	121 (49·8%)	123 (50·4%)	121 (49·8%)	365 (50·0%)
North America	56 (23·0%)	55 (22·5%)	56 (23·0%)	167 (22·9%)
Others	31 (12·8%)	32 (13·1%)	32 (13·2%)	95 (13·0%)

Example: For the Placebo

```
   n=243
X.sex=sample( c(rep("Female",181),rep("Male",n-181)))
X.region=sample( c(rep("Asia",35),rep("Europe",121),
rep("North␣America",56),rep("Others",31))     )
```

- Generate continuous covariates: Baseline. The log transform of the 24-h cough frequency follows the normal distribution with variance 1 and mean providing in the following table.



Baseline values for efficacy endpoints				
24-h cough frequency				
Geometric mean coughs per h	22·0	19·7	17·7	19·7

Example: For the Placebo

```
baseline=sapply( rnorm(n, log(22),1),function(x) max(x,0) )
```

I feel weird that it did not provide the standard deviation of the baseline. Here are the coefficients

- Generate response Y_{ij} $j = 1, \ldots, m$ (24-h cough frequency at the $j-$th visit.)

$$log(Y_{ij}) = log(baseline) + Treatment + Sex + Region + visit + \epsilon_{ij} \tag{7}$$

and the noise term $\boldsymbol{\epsilon}_i$ follows the multivariate normal distribution generated by the following code.

```
        mvrnorm(n,mu=rep(0,m),sigma=0.5*(0.4*diag(m)+0.6))
```

- Generate binary response. If the 24-h cough frequency reduces more than x%, the binary response is one, otherwise it is zero. x can be 30.

$$\begin{aligned} binary.Y_{ij} &= 1 \quad \text{if } Y_{ij}/baseline < (1 - x\%) \\ &= 0 \quad \text{other.} \end{aligned}$$

Figure 3 and figure 4 show the distribution of the 24-hour cough frequency of the simulated clinical trial and the actual data, respectively. We can see that the simulated clinical trial data is closed to the real data.

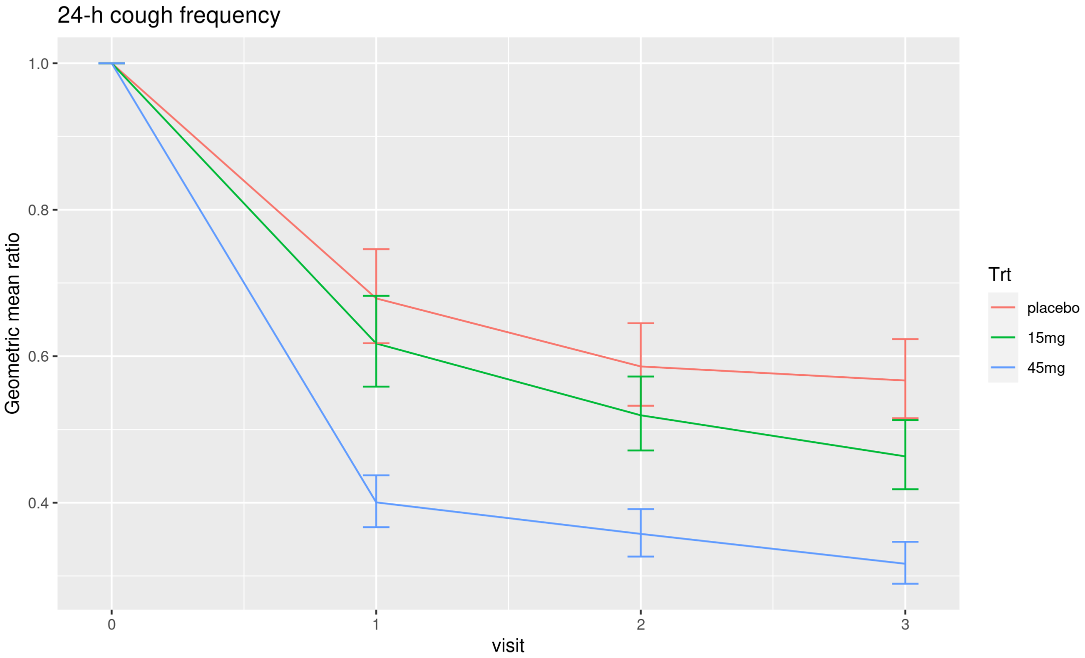


Figure 3: 24-h cough frequency from simulated clinical trial data, the error bars is 95% confidence intervals.

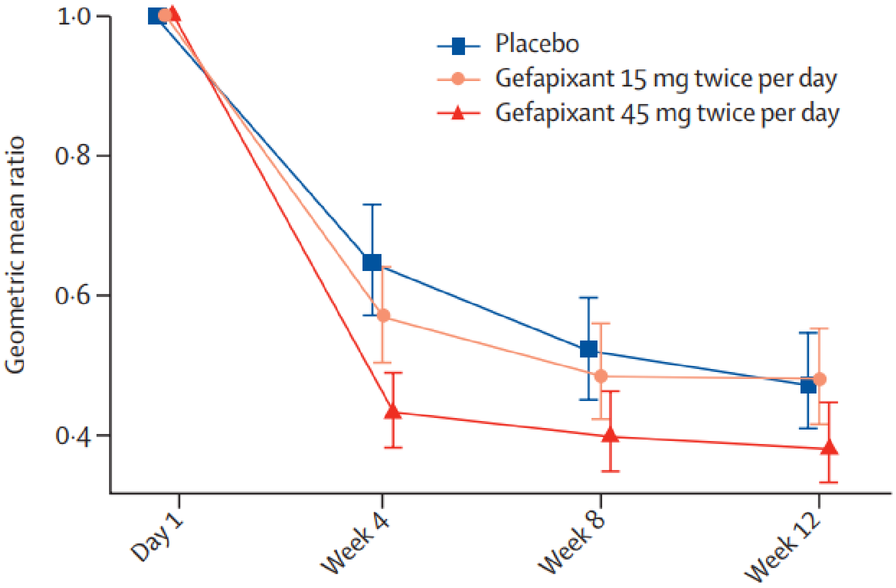


Figure 4: 24-h cough frequency from real data,the error bars is 95% confidence intervals.

4.1 Cross-sectional Study

Although the simulated data is longitudinal, we can focus at a single time point and treat is as a cross-sectional study. We consider the following working models

$$logP(Y=1|\boldsymbol{X}) \sim Trt + Sex + Region + baseline + Region * baseline + baseline^2.$$

From table 3, there is no significant difference between the two methods. The simulation in Section 2.2 shows that the performance of LBC and Poisson are similar in a large sample size. The sample size of this study is 730, which is larger than the one in the simulation. And hence, that can be a reason that we did not observe any difference. Table 4 shows the convergence rates. One can see that even if the sample size is large as 730, log-binomial models may not converge. And hence, it is crucial to add constraints.

		Visit1		Visit2		Visit3	
Trt	Measure	LBC	Poisson	LBC	Poisson	LBC	Poisson
15mg	Coverage	0.95	0.95	0.94	0.94	0.95	0.95
	length	0.38	0.38	0.34	0.34	0.31	0.31
45mg	Coverage	0.95	0.96	0.94	0.95	0.93	0.95
	length	0.32	0.32	0.29	0.29	0.26	0.26

Table 3: 95% confidence interval. The success rate is around 50% $\sim$ 70%.

Method	visit1	visit2	visit3
LB	0.96	0.88	0.67
LBC	1	1	1
Poisson	1	1	1

Table 4: The convergence rates

4.2 Longitudinal Study

In this section, we use the whole data set and compare the performance between LBC-GEE and Poisson-GEE. We consider a working models with interaction term between "visit" and "treatment". Please see the document "Simulated-Clinical-Trial.html" showing the detail of estimating the risk ratio at a single time point.

$$logP(Y=1) \sim Trt + Sex + Region + baseline + visit + Trt * visit + baseline * visit. \tag{8}$$

Table 5 shows the bias and the 95% confidence intervals for the log risk ratio at a single time point. Unfortunately, there is a considerable bias for LBC-GEE for the treatment "45mg". As a result, the coverage rate is not satisfactory. We know that there is model misspecification for simulated clinical trial data. In the working model (8), we implicitly assume that the risk ratio is constant for every patient. However, from Figure 5, which shows the distribution of the true risk ratio of each patient, we can see that this assumption is not valid for "45mg".

		visit1		visit2		visit3	
Trt		LBC	Poi	LBC	Poi	LBC	Poi
15mg	Bias	0.01	0.01	0.01	0.01	0.01	0.02
	Coverage	0.93	0.95	0.94	0.94	0.93	0.94
45mg	Bias	-0.00	0.01	-0.02	0.00	-0.03	0.00
	Coverage	0.93	0.96	0.92	0.93	0.90	0.94

Table 5: 95% confidence intervals and bias. The number of replications is 400.

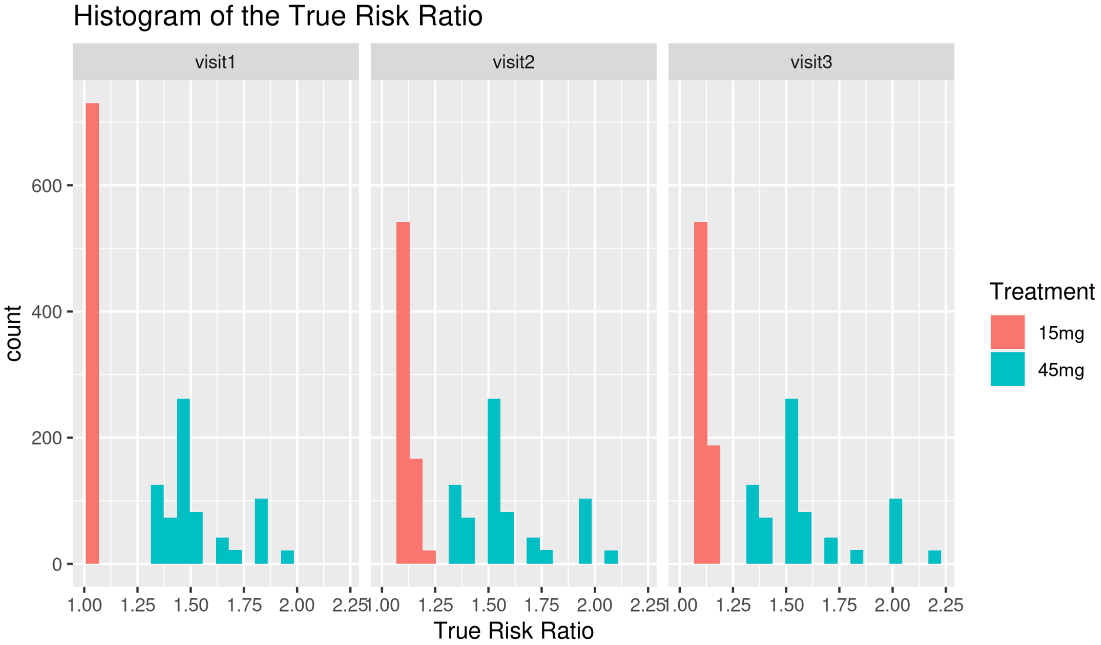


Figure 5: Distribution of True Risk Ratio

5 Summary and discussion

Interpretation is always a crucial goal for statisticians. Summarizing the binary data by risk ratio can be easier to explain and understand by our stakeholders. The log-binomial model coefficient indicates the log risk ratio, which can be a candidate for modeling the binary data. The log-binomial models have several benefits. First, it can provide a probability estimate. So, the log-binomial model makes more sense to the scientist. Second, it can reach the Cramér–Rao lower bound. So, it is optimal. However, the log-binomial models may not converge, especially in the presence of multi-dimension and longitudinal data. In this paper, we show that the constraints can solve the convergence issue of the log-binomial models and compare the performance of the log-binomial models with constraints to another well-known algorithm, Poisson regression.

5.1 Cross-sectional Study

We first show that the log-binomial models in the cross-sectional study may not converge. Especially when the dimension p is large or the sample size, n is small. And the LBC can solve the convergence issue for LB. LBC has a better coverage rate for confidence intervals than Poisson regression when the dimension p is large, the success rate $mean(Y)$ is small, or the sample size is small. And the other cases, LBC and Poisson are similar to each other. Furthermore, since the sample size of the simulated clinical trial data is relatively large, the performance of LBC is similar to Poisson.

5.2 Longitudinal Study

Applying the log-binomial model to a longitudinal is still an open problem. This paper proposes a recursive algorithm inspired by generalized estimating equations. We answer three questions for our proposed algorithm (LBC-GEE). How to add constraints? How to estimate correlation structure for binary data? And how to stop this recursive algorithm? LBC-GEE can perfectly solve the convergence issue. And, the mean square error of LBC-GEE can have an improvement of up to 4 % compared to the GEE type of Poisson regression. However, in the simulation studies, we face several challenges. First, although LBC-GEE can converge and be consistent, the confidence intervals are not valid when the success rate $mean(Y)$ is large. And hence, we need to find a better way to estimate the variance. Bootstrapping can be a possible solution, but we need to speed up the algorithm before applying the bootstrapping. Now, one simulation for LBC-GEE takes around 1 min. Second, in the simulated clinical trial data, we realized there is a considerable bias while the constant risk ratio assumption is incorrect.

6 Supplementary

- Document ”LBC.html” shows how to implement the log-binomial model with constraints.
- Document ”LBC-GEE.html” shows how to implement our proposed algorithm LBC-GEE for longitudinal data and how to generate simulation data in Section 3.4.

- Document ”Simulated-Clinical-Trial.html” shows how to generate simulated clinical data and how we apply LBC-GEE to the data.

References

de Andrade, B. B. & Andrade, J. M. d. L. (2018), ‘Some results for maximum likelihood estimation of adjusted relative risks’, *Communications in Statistics-Theory and Methods* **47**(23), 5750–5769.

Luo, J., Zhang, J. & Sun, H. (2014), ‘Estimation of relative risk using a log-binomial model with constraints’, *Computational Statistics* **29**(5), 981–1003.

McGarvey, L. P., Birring, S. S., Morice, A. H., Dicpinigaitis, P. V., Pavord, I. D., Schelfhout, J., Nguyen, A. M., Li, Q., Tzontcheva, A., Iskold, B. et al. (2022), ‘Efficacy and safety of gefapixant, a p2x3 receptor antagonist, in refractory chronic cough and unexplained chronic cough (cough-1 and cough-2): results from two double-blind, randomised, parallel-group, placebo-controlled, phase 3 trials’, *The Lancet* **399**(10328), 909–923.

Rao Chaganty, N. & Joe, H. (2004), ‘Efficiency of generalized estimating equations for binary responses’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(4), 851–860.

Schwendinger, F., Grün, B. & Hornik, K. (2021), ‘A comparison of optimization solvers for log binomial regression including conic programming’, *Computational Statistics* **36**(3), 1721–1754.

Wang, M. (2014), ‘Generalized estimating equations in longitudinal data analysis: a review and recent developments’, *Advances in Statistics* **2014**.

Williamson, T., Eliasziw, M. & Fick, G. H. (2013), ‘Log-binomial models: exploring failed convergence’, *Emerging themes in epidemiology* **10**(1), 1–10.

Zou, G. (2004), ‘A modified poisson regression approach to prospective studies with binary data’, *American journal of epidemiology* **159**(7), 702–706.