Log-binomial Regression Models for HTA Submissions

Chi-Shian Dai

BARDS 2022 Summer Intern
University of Wisconsin-Madison

08/24/2022

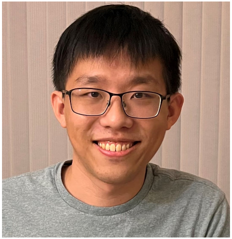
Team

William Malbecq

Shahrul Mt-Isa

Richard Baumgartner

Chi-Shian Dai



Motivation

- Binary Response.
- Goal: Risk ratio, not odds ratio.

	Risk Ratio	Odds Ratio
Model	log-binomial	logistic
Interpretation	✓	X

- Challenge: The log-binomial regression may not converge, especially the dimension is large.
- A simple log-binomial model with a single covariate (treatment) may not be problematic.
- Cross-sectional, and longitudinal study.
- Extreme cases: Dimension is large, and the success rate is low or large.

Agenda

1. Cross-sectional Study
2. Longitudinal Study
3. Summary

Agenda

Log-Binomial Models (LB)

- Binary Response: y_i for patient $i = 1, \ldots, n$.
- Covariates: $\boldsymbol{x}_i$ for patient $i = 1, \ldots, n$.
- Treatment: Trt_i for patient $i = 1, \ldots, n$.
- Models: Log link function.

$$\log P(Y_i = 1 | \boldsymbol{x}_i, Trt) = \alpha Trt + \boldsymbol{x}_i^\top \beta$$

- Risk Ratio:

$$P(Y_i = 1 | \boldsymbol{x}_i, Trt = 1) / P(Y_i = 1 | \boldsymbol{x}_i, Trt = 0) = e^{\alpha}$$

- MLE:

$$\widehat{\gamma} = \arg \max_{\gamma = (\beta, \alpha)} \ell(\alpha, \beta)$$

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

- It may not converge.

Log-binomial Models with Constraints (LBC)

- Add constraints. [3]

$$\widehat{\gamma} = \arg \max_{\gamma=(\beta,\alpha)} \ell(\alpha, \beta)$$

$$\text{Subject to} \quad \alpha Trt_i + \boldsymbol{x}_i^\top \beta < 0 \quad \forall i = 1, \ldots, n$$
$$\equiv P(Y_i = 1|\alpha, \beta) < 1$$

- Conic Programming : Use **ROI** package in **R**.[9].

$$\arg \min_{\boldsymbol{x}} \boldsymbol{x}^\top a$$
$$\text{subject to } b - A\boldsymbol{x} \in \mathbf{K},$$

where **K** is a cone.

Adjusted Confidence Intervals

1. Calculate $\widehat{\gamma} = (\widehat{\alpha}, \widehat{\beta})$, and the Fisher information matrix $I(\gamma)$.
2. Let A be the matrix that collects the rows of the designed matrix X satisfying $\widehat{\alpha} Trt_i + \boldsymbol{x}_i^\top \widehat{\beta} = 0$. (The designed matrix includes both covariates $\boldsymbol{x}$ and treatment Trt.)
3. We have the following asymptotic theory [1].

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

where

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

Poisson Regression for Risk Ratio

- Ignore that it is binary responses, then apply the Poisson regression with a log link function.
- Pros:
  1. It converges. (No boundary issue.)
  2. It is easy to implement.
  3. It is consistent.
- Cons:
  1. The estimated probability can be greater than one.
  2. It does not approach Cramér–Rao lower bound.
- Confidence Intervals: Sandwich method.

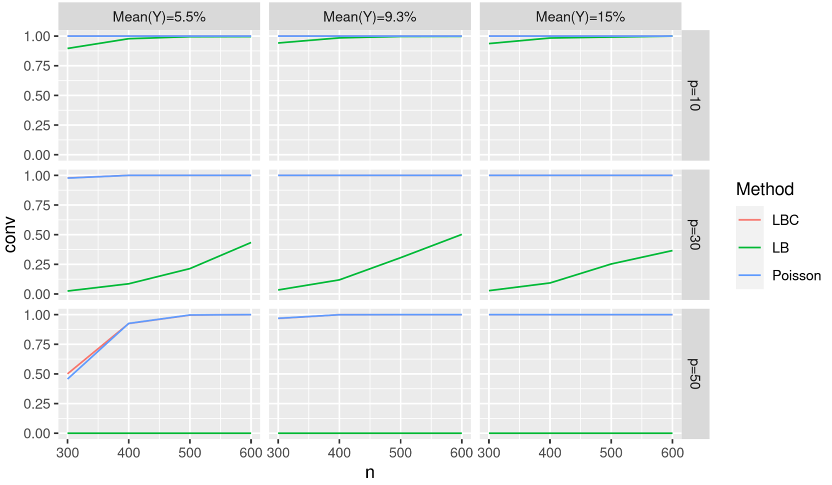
Simulation

- Covariates: $\boldsymbol{X} = (X_1, \ldots, X_p)$ $i.i.d.$ $Unif(0,1)$.
- Treatment: Randomized controlled trial with a probability of 0.5.
- Response:

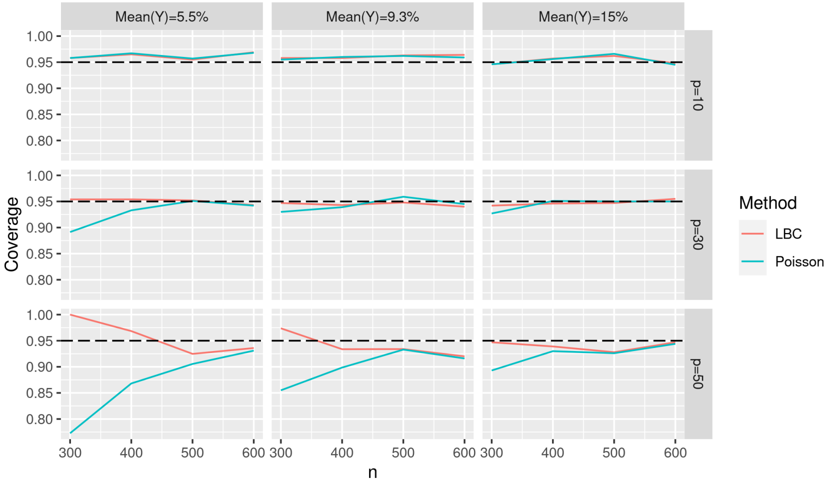
$$\log P(Y = 1|\boldsymbol{X}, Trt) = \log(3)\,Trt + c_0 - \sum_{j=1}^{p} 0.5^p X_j$$

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

Convergence Rates



Coverage Probability of 95% Confidence Intervals



Simulated Clinical trial Study

- Clinical trial[4]: To confirm the efficacy of gefapixant for chronic cough.
- A longitudinal study with 3 visits and 730 patients.
- Response: 24-h cough frequency
  Dichotomization: 30% reduction or more in 24-h cough frequency.
- Treatments: Placebo, 15mg, 45mg. (1:1:1)
- Covariates: Sex, Region, Baseline(24-h cough frequency in the first day.)
- Working Models: Look at a single time point. (Cross-sectional Study)

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

The dimension $p = 11$.

Simulated Clinical trial Study

1. Generate covariate (Sex, Region, Baseline) follows the marginal distribution given by [4].
2. Generate response Y_{ij} $j = 1, \ldots, 3$ with the following models and given coefficients.

$$log(Y_{ij}) = log(baseline) + Treatment + Sex + Region + visit + \epsilon_{ij}, \quad (2)$$

   The noise term ϵ_i follows the multivariate normal distribution with mean 0, variance 0.5 and correlation 0.6.
3. Dichotomization: 30% reduction or more in 24-h cough frequency.

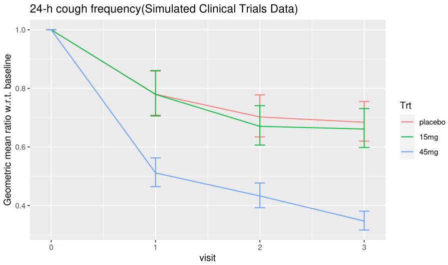


Figure: 24-h cough frequency from simulated data, the error bars are 95% confidence intervals.

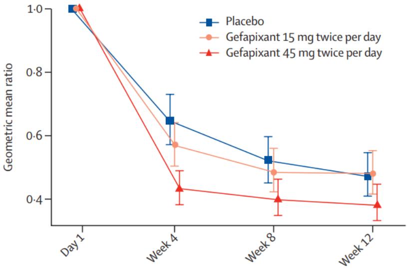


Figure: 24-h cough frequency from real data, the error bars are 95% confidence intervals.

Simulated Clinical trial Study

		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: 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: The convergence rates

Comparison between LBC and Poisson

	LBC	Poisson
Convergence	✓	✓
Probability	✓	X
Cramér–Rao bound	✓	X
Easy Implement	X	✓
Time	Slow	Fast

Summary for Cross-sectional Study

- Log-binomial models provide risk ratio.
- Log-binomial models may not converge. Especially when the dimension p is large or the sample size, n is small.
- Adding constraints is really helpful for the convergence issue.
- Adjusted confidence Intervals for LBC.
- 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.
- The performances of LBC and Poisson are similar when the dimension p is small, the success rate $mean(Y)$ is large, or the sample size is large.
- In the simulated clinical trial data, LBC is similar to Poisson.

MSD

Agenda

Log-Binomial models for Longitudinal Data

- For each patient $i = 1, \ldots n$.
  Repeated measurements: y_{ij}, $j = 1, \ldots m$.
- Covariates: $\boldsymbol{x}_{ij}$ with dimension p.
- Generalize estimating equations (GEE) type of log-binomial models.
- Problems with the existing packages. EX: **gee** and **geepack** in **R**.
  1. Converge rate is around 1% in the simulation.
  2. The correlation structure is not appropriate for binary response data.

Review: Generalize estimating equations

- Find $\widehat{\boldsymbol{\beta}}$ solve the following equations.

$$\sum_{i}^{n} \nabla \boldsymbol{\mu}_i(\boldsymbol{\beta})^\top V_i^{-1}(\boldsymbol{\beta})\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = 0, \tag{3}$$

- $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{im}(\boldsymbol{\beta}))$ is the estimator for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})$,
- $\mu_{ij}(\boldsymbol{\beta}) = P(y_{ij} = 1 | \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) = \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})$,
- V_i is the variance of $\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})$

Questions:

1. How to add constraints to (3)?
2. How to estimate V_i.

GEE-types of LBC

- If we have constant estimates $\widehat{V}_i$ for $V_i(\boldsymbol{\beta}^*)$, we have the following constrained GEE

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

1. Ignore the longitudinal structure. Get $\widehat{\boldsymbol{\beta}}_0$ by the standard LBC.
2. Use $\widehat{\boldsymbol{\beta}}_k$ to estimate $\widehat{V}_i$.
3. Get the new estimate $\widehat{\boldsymbol{\beta}}_{k+1}$ from the above constrained GEE.
4. Repeat step 2 and step 3 until $\widehat{\boldsymbol{\beta}}_k$ reaches the stopping rule.

How to estimate V_i?

- To get $\widehat{V}_i$, we need to estimate the correlation structure $Cor(y_{ij}, y_{ik}|\widehat{\beta})$.
- In GEE, people consider a constant correlation structure. That means $Cor(y_{ij}, y_{ik}|\widehat{\beta})$ is constant w.r.t. patient i.
  Ex: package **gee** and **geepack** in **R**.
- The constant correlation structure is not appropriate for binary response since the domain for $Cor(y_{ij}, y_{ik}|\widehat{\beta})$ is not (-1,1).

$$-\sqrt{\frac{P_j P_k}{(1-P_j)(1-P_k)}} \leq Cor(y_{ij}, y_{ik}) \leq \frac{\min\{P_j, P_k\} - P_j P_k}{\sqrt{P_j P_k (1-P_j)(1-P_k)}},$$

- Estimate $p(y_{ij} y_{ik} = 1|\widehat{\beta})$.

Estimate Joint distribution

- Assume that the binary responses $(Y_{i1}, \dots Y_{im})$ comes from dichotomizing a multivariate normal distribution with an exchangeable correlation.
- We have the estimated marginal probability $\widehat{p}_{ij}$.

1. Find the normal quantiles q_{ij}.

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

2. 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.

3. Find MLE for ρ.

Adjusted Confidence Intervals:

1. Calculate sandwich covariance Σ.
2. Find the adjusted covariance by substitute sandwich covariance with I in (1).

$$\Sigma_{adjust} = \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma.$$

Simulation for Longitudinal Data

- Methodologies:
  - GEE type of log-binomial with constraints(LBC): We have a function call "fit.lbc.gee".
  - GEE type of Poisson regression: **geeglm** function in the package **geepack** with an exchangeable correlation.
- Covariates: $\boldsymbol{X} = (X_1, \ldots, X_p)$ $i.i.d.$ $Unif(0, 1)$.
- Treatment:

$$T_k \sim Ber(0.5)$$

Simulation for Longitudinal Data

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

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

$$\begin{aligned} \log P_{ij} &= \log p(Y_{ij} = 1 | T_i, \boldsymbol{X}_i) \\ &= aT_i - \log(3) - c - X_{i1} - 0.5X_{i2}, \end{aligned}$$

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, $Cov(Z_i, Z_j) = 0.6$.
4. Let $Y_{ij} = 1$ if $Z_{ij} \leq q_{ij}$, otherwise $Y_{ij} = 0$.

95% confidence intervals

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

Table: 95% confidence intervals, $p = 2,\ m = 5$.

Relative Mean Square Error

- Mean Square Error(MSE):

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

- Relative Mean Square Error(%):

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

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

Table: Relative Mean Square Error (%)

Simulated Clinical trial Study: Confidence Intervals

- Working Models:

$$\begin{aligned} logP(Y=1) \sim & Trt + Sex + Region + baseline + visit \\ & + Trt * visit + baseline * visit. \end{aligned}$$

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

Table: 95% confidence intervals

MSD

Summary

Longitudinal Study

- Develop a recursive algorithm for the GEE-type log-binomial models.
- It solves the convergence issue of the existing packages.
- We consider a non-constant correlation structure.
- In the simulation, LBC-GEE, and Poisson-GEE are consistent. However, in some cases, the coverage rates of LBC-GEE are not satisfactory. We need to find a better way to estimate the variance.
- LBC-GEE has a smaller mean square error than Poisson-GEE.
- In the simulated clinical trial study, LBC-GEE has a considerable bias.

Summary

Cross-Sectional Study

- Log-binomial models may not converge. Especially when the dimension p is large or the sample size, n is small.
- Adding constraints is really helpful for the convergence issue.
- Adjusted confidence Intervals for LBC.
- 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.

Thank you

References I

[1] Bernardo Borba de Andrade and Joanlise Marco de Leon Andrade. "Some results for maximum likelihood estimation of adjusted relative risks". In: *Communications in Statistics-Theory and Methods* 47.23 (2018), pp. 5750–5769.

[2] Yihan Li et al. "Analyzing longitudinal binary data in clinical studies". In: *Contemporary Clinical Trials* (2022), p. 106717.

[3] Ji Luo, Jiajia Zhang, and Han Sun. "Estimation of relative risk using a log-binomial model with constraints". In: *Computational Statistics* 29.5 (2014), pp. 981–1003.

[4] Lorcan P McGarvey et al. "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". In: *The Lancet* 399.10328 (2022), pp. 909–923.

References II

[5] Myunghee Cho Paik. "The generalized estimating equation approach when data are not missing completely at random". In: *Journal of the American Statistical Association* 92.440 (1997), pp. 1320–1329.

[6] N Rao Chaganty and Harry Joe. "Efficiency of generalized estimating equations for binary responses". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.4 (2004), pp. 851–860.

[7] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data". In: *Journal of the american statistical association* 90.429 (1995), pp. 106–121.

References III

[8] Florian Schwendinger, Bettina Grün, and Kurt Hornik. "A comparison of optimization solvers for log binomial regression including conic programming". In: *Computational Statistics* 36.3 (2021), pp. 1721–1754.

[9] Stefan Theußl, Florian Schwendinger, and Kurt Hornik. "ROI: an extensible R optimization infrastructure". In: (2019).

[10] Ming Wang. "Generalized estimating equations in longitudinal data analysis: a review and recent developments". In: *Advances in Statistics* 2014 (2014).

[11] Fang Xie and Myunghee Cho Paik. "Generalized estimating equation model for binary outcomes with missing covariates". In: *Biometrics* (1997), pp. 1458–1466.

Stopping Rule for LBC-GEE

- In theory, $\widehat{\boldsymbol{\beta}}_k$ is consistent for every k.
- 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).

n	Case	%
200	No-Stop	0.03
	Converge	0.96
	Early-stop	0.01
300	No-Stop	0.01
	Converge	0.98
	Early-stop	0.01
600	No-Stop	0.01
	Converge	0.97
	Early-stop	0.02

Table: A simulation with $p = 10,\ m = 5$.

95% Confidence Intervals lengh

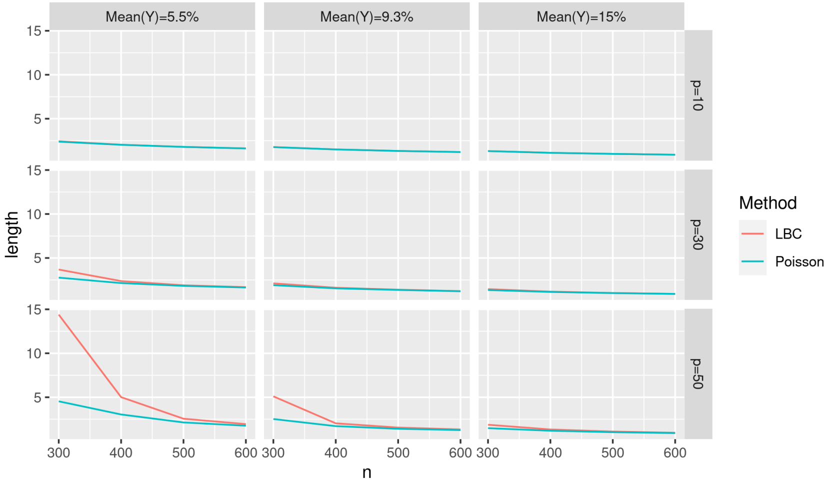


Figure: Length of 95% confidence intervals.

ROI package

- ROI: R Optimization Infrastructure
- Provides an extensible infrastructure to solve optimization problems in a consistent way.
- User: Can easily apply different solvers.

Back to the current slide.

Conic Programming

- Conic:

$$\arg\min_{\boldsymbol{x}} \boldsymbol{x}^\top a$$
$$\text{subject to } b - A\boldsymbol{x} \in \mathbf{K},$$

where $\mathbf{K}$ is a cone.

Back to the current slide.

Sandwich Method

- It is a way to estimate standard deviation for the solution of estimating equations.
- It is consistent as long as the estimating equation is correct.
- Poisson Regression for Risk Ratio: The model is not correct, but the estimating equation(a derivative of the log-likelihood function) is correct.

Back to the current slide.

Distribution of True Risk Ratio

Boxplots of the Bias