Bootstrap Estimation of Disease Incidence Proportion with Measurement Errors

Masataka Taguri (Univ. of Tokyo.)

Hisayuki Tsukuma (Toho Univ.)

Contents

- Motivation
- Formulation of the Problem
- Regression Calibration Method
- Proposed Estimator
- Interval Estimation (Delta & Bootstrap)
- Numerical Examination

Motivation

- Measurement error problem
  - The measurement error of exposure variable will tend to give bias estimates of relative risk
    - ⟨ex⟩ In nutritional studies, FFQ (food frequency questionnaire) is not exact measurement

★ How to obtain correct risk estimates?

Formulation of Problem (1)

- Data form
  - Main data : (Y_i, Q_i) $(i = 1,2,\cdots,n)$

    Y_i : binary response (Disease or not)

    Q_i : exposure measured with error

    n : No. of main data-set
  - Validation data : (T_j, Q_j) $(j = 1,2,\cdots,m)$

    T_j : true exposure

    m : No. of validation data-set

Formulation of Problem (2)

- Disease model

$$\text{logit}[\Pr(D|T)] = \alpha_0 + \alpha_1 T$$

- Assume linear relationship of T and Q

$$T = \lambda_0 + \lambda_1 Q + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

- Naive model

$$\text{logit}[\Pr(D|Q)] = \gamma_0 + \gamma_1 Q \quad (Q: \text{observable})$$

★ Usual interest : estimation of α_1

Regression Calibration (RC) method

- Model

$$\text{logit}[\Pr(D|T)] = \alpha_0 + \alpha_1 T$$

$$T = \lambda_0 + \lambda_1 Q + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\text{logit}[\Pr(D|Q)] = \gamma_0 + \gamma_1 Q$$

(i) estimate γ_1 by maximum likelihood

(ii) estimate λ_1 by ordinary least squares

(iii) estimate α_1 by $\hat{\alpha}_1 = \hat{\gamma}_1 / \hat{\lambda}_1$

Rosner, et al.1989

Asymptotic Variance of $\hat{\alpha}_1$

- Apply the delta method, then asymptotic variance of α_1 is given by

$$Var\ (\hat{\alpha}_1) = Var\left(\frac{\hat{\gamma}_1}{\hat{\lambda}_1}\right)$$

$$\approx \left(\begin{array}{c} \frac{\partial \alpha_1}{\partial \gamma_1} \\ \frac{\partial \alpha_1}{\partial \lambda_1} \end{array}\right)^t_{\theta=\hat{\theta}} \begin{pmatrix} Var\ (\hat{\gamma}_1) & 0 \\ 0 & Var\left(\hat{\lambda}_1\right) \end{pmatrix} \left(\begin{array}{c} \frac{\partial \alpha_1}{\partial \gamma_1} \\ \frac{\partial \alpha_1}{\partial \lambda_1} \end{array}\right)_{\theta=\hat{\theta}}$$

$$= \frac{1}{\hat{\lambda}_1^2} Var\ (\hat{\gamma}_1) + \frac{\hat{\gamma}_1^{\,2}}{\hat{\lambda}_1^4} Var\left(\hat{\lambda}_1\right)$$

Subject for investigation

★ RC estimator $\hat{\alpha}_1 = \frac{\hat{\gamma}_1}{\hat{\lambda}_1}$: ($\hat{\gamma}_1$ ← MLE, $\hat{\lambda}_1$ ← OLSE)

might be unstable (heavy tailed)

- $\hat{\gamma}_1, \hat{\lambda}_1$: asymptotically ratio of two normal variables

Objective of our study

① To propose an estimator of α_1
- examination of the stability of estimators

② Interval estimation of α_1
- Bootstrap Method [Percentile & BCa]
- Delta Method

③ Interval estimation of $p = \Pr(D|T)$
- Bootstrap Method [Percentile & BCa]
- Delta Method

Inverse Regression Estimator

- To construct alternative estimator of α_1, consider the inverse regression model;

$$Q_j = \beta_0 + \beta_1 T_j + \varepsilon_j$$

- The ordinary least squares estimator of β_1

$$\hat{\beta}_1 = \hat{\lambda}_1 \Big/ \left\{ \left(S_e / S_{QQ} \right) + \hat{\lambda}_1^2 \right\}$$

$$S_e = \sum_{j=1}^{m} (T_j - \hat{\lambda}_0 - \hat{\lambda}_1 Q_j)^2 \quad S_{QQ} = \sum_{j=1}^{m} (Q_j - \overline{Q})^2$$

Generalized Estimator

- Ridge-type Estimator for $1/\lambda_1$

$$\hat{\beta}_{1k} = \hat{\lambda}_1 \Big/ \left\{ \left(k\hat{\sigma}^2 \Big/ S_{QQ} \right) + \hat{\lambda}_1^2 \right\}, \quad \hat{\sigma}^2 = S_e / m-2$$

- Generalized Inverse Regression Estimator

$$\hat{\alpha}_{1k} = \hat{\gamma}_1 \hat{\beta}_{1k} = \hat{\gamma}_1 \times \frac{\hat{\lambda}_1}{k\hat{\sigma}^2 \Big/ S_{QQ} + \hat{\lambda}_1^2}$$

★ $k = 0 \Rightarrow$ RC estimator

★ The asymptotic distribution of $\hat{\beta}_1$ has moments.

Interval Estimation of α_1 (by Delta method)

[Algorithm]

1) Use $\hat{\alpha}_1$ and $\hat{\alpha}_{1k}$ to construct the confidence intervals of α_1
2) Estimate the asymptotic variance of $\hat{\alpha}_1$ and $\hat{\alpha}_{1k}$ by delta method
3) Make confidence intervals by normal approximation

Rosner, et al.1989

Interval Estimation of α_1 (by Bootstrap method)

- Estimator : $\hat{\alpha}_1(\hat{\alpha}_{1k})$

[Algorithm]

1) Compute $\hat{\gamma}_1^{*b}(b=1,2,\cdots,B)$ by resamples from main data-set.
2) Compute $\hat{\lambda}_1^{*b}(b=1,2,\cdots,B)$ by resamples from validation data-set.
3) Compute $\hat{\alpha}_1^{*b} = \hat{\gamma}_1^{*b} / \hat{\lambda}_1^{*b}$ for $b=1,2,\cdots,B$.
4) Construct confidence intervals of α_1 by Bootstrap Method [Percentile & BCa]. (Efron & Tibshirani, 1993)

Interval Estimation of $p = \Pr(D|T)$

- Estimator of *p* :

$$\hat{p} = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 T)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 T)}$$

$\hat{\alpha}_0, \hat{\alpha}_1$: RC / Inverse Regression Est.

- Confidence Interval (C.I.) of *p* :

  by Delta or Bootstrap (Percentile & BCa)

★ *p* should exist between 0 and 1

⇒ C.I. of logit(*p*) → C.I. of *p*

Numerical Examination – Set-up

- Validation sample

1° Generate T_i from $N(0,1)$ $(i = 1,2,\cdots,m).$

2° Assume the model

$$Q_i = T_i + e_i, \quad e_i \sim N(0, (1-\lambda_1)/\lambda_1).$$

Generate e_i $(i = 1,2,\cdots,m).$

Make validation sample (T_i, Q_i) $(i = 1,2,\cdots,m)$

for $\lambda_1 = 0.3,\ 0.5,\ 0.7.$

★ The above model is a special case of

$$T = \lambda_0 + \lambda_1 Q + \varepsilon,\ \varepsilon \sim N(0,\sigma^2) \quad [\lambda_0 = 0,\ \sigma^2 = 1-\lambda_1]$$

- Main sample

3° Generate (T_i, Q_i) $(i = 1,2,\cdots,n)$.

$$\exp(\alpha_1) = 1.5, 2.0, 3.0, 4.0, 5.0\,; \quad e^{\alpha_0}/(1+e^{\alpha_0}) = 0.05$$

$$\Rightarrow \quad p_i = \exp(\alpha_0 + \alpha_1 T_i)/\{1+\exp(\alpha_0 + \alpha_1 T_i)\}$$

Generate Y_i from Bernoulli dist.(p_i).

Combine Y_i's with Q_i's (paired value of T_i's)

$\Rightarrow$ $(Y_i,\ Q_i)$ $(i = 1,2,\cdots,n)$.)

4° Set $T = -1,\ 0,\ 1,\ 2$.

Compute 95% C.I. of logit(p), then p

by Normal approximation, Percentile & BCa.

★ Set $m = 100,\ n = 1000,\ R = 1000$

▪ Computation : for all combination of $\lambda_1,\ \alpha_1$.

Table 1 Estimates of α_1

(R=1000, B=2000, n=1000, m=100)

Esti-mator	Mean	SD	Skew-ness	Kurto-sis	Min	Max	Range
α_1=0.4050 (exp(α_1)=1.5), λ_1=0.5							
RC → α_1^R	0.409	0.214	0.227	3.298	-0.32	1.249	1.573
GI → α_1^G	0.405	0.212	0.224	3.297	-0.32	1.238	1.556
α_1=0.6931 (exp(α_1)=2.0), λ_1=0.5							
α_1^R	0.692	0.217	0.201	3.220	0.009	1.528	1.519
α_1^G	0.685	0.214	0.191	3.208	0.009	1.503	1.494
α_1=1.0986 (exp(α_1)=3.0), λ_1=0.5							
α_1^R	1.052	0.217	0.413	3.235	0.516	1.924	1.408
α_1^G	1.041	0.213	0.402	3.213	0.512	1.892	1.380

Examination on stability of α_1

(1) $\hat{\alpha}_1^G$ underestimates except $e^{\alpha_1} = 3.0,\ \lambda_1 = 0.7.$

$\hat{\alpha}_1^G,\ \hat{\alpha}_1^R$: monotone increasing w.r.t. λ_1

except $e^{\alpha_1} = 1.5.$ for $\hat{\alpha}_1^R$.

$\hat{\alpha}_1^G$ is smaller than $\hat{\alpha}_1^R$ for all cases.

Biases of $\hat{\alpha}_1^G$, $\hat{\alpha}_1^R$: not so different on the whole.

(2) SD, skewness, kurtosis, range :

values for $\hat{\alpha}_1^R$ **>** values for $\hat{\alpha}_1^G$

$\Rightarrow$ Tail of $\hat{\alpha}_1^R$'s distribution is heavier than that of $\hat{\alpha}_1^G$'s $\rightarrow$ Outliers may often appear.

Table 2-1 95% Confidence Intervals of α_1

(R=1000, B=2000, n=1000, m=100)

α_1=0.4050 (exp(α_1)=1.5), λ_1=0.5			
Method	Cov. Prob.	Length	Shape
Nor.app. NR	0.956	0.8212	1.0000
NG	0.954	0.8116	1.0000
Parcentile PR	0.942	0.8495	1.1323
PG	0.945	0.8377	1.1265
BCa BR	0.958	0.8846	0.6939
BG	0.956	0.8729	0.6920

Table 2-2 95% Confidence Intervals of α_1
（R=1000, B=2000, n=1000, m=100)

α_1＝0.6931 (exp(α_1)=2.0), λ_1＝0.5			
Method	Cov. Prob.	Length	Shape
NR	0.945	0.8214	1.0000
NG	0.943	0.8107	1.0000
PR	0.935	0.8516	1.2311
PG	0.935	0.8378	1.2209
BR	0.930	0.8699	0.7114
BG	0.927	0.8575	0.7082

Table 2-3 95% Confidence Intervals of α_1
（R=1000, B=2000, n=1000, m=100)

α_1＝1.0986 (exp(α_1)=3.0), λ_1＝0.5			
Method	Cov. Prob.	Length	Shape
NR	0.930	0.8376	1.0000
NG	0.922	0.8246	1.0000
PR	0.937	0.8722	1.3589
PG	0.935	0.8543	1.3431
BR	0.896	0.8683	0.7664
BG	0.890	0.8534	0.7605

Examination on confidence interval of α_1

(1) Cov. Prob. of NG **<** Cov. Prob. of NR.

Cov. Prob. of normal approximation :

monotone decreasing w.r.t. e^{α_1}(odds ratio)

(2) Length of C.I. by $\hat{\alpha}_1^G$ **<** Length of C.I. by $\hat{\alpha}_1^R$

(for all cases) ⇒ stable property of $\hat{\alpha}_1^G$

(3) Bootstrap : slightly worse than Normal approxim.

（from viewpoint of Cov. Prob.）

★ Percentile is best in case of $e^{\alpha_1} = 3.0$.

(4) BCa：not better than other methods

⇔ The distribution of estimates is not so skew?

Table 3-1 95% Confidence Intervals of p

(R=1000, B=2000, n=1000, m=100 ; $\exp(\alpha_1)$=2.0, λ_1=0.5)

Method	Cov. Prob.	Length	Shape
T=-1, Pr[D|T]=0.0256			
NR	0.917	0.0342	1.7567
NG	0.919	0.0341	1.7481
PR	0.933	0.0318	1.2882
PG	0.933	0.0321	1.3169
BR	0.954	0.0314	0.8901
BG	0.954	0.0315	0.9193
T=0, Pr[D|T]=0.05			
NR	0.905	0.0333	1.3310
NG	0.905	0.0332	1.3301
PR	0.919	0.0313	1.0245
PG	0.927	0.0314	1.0585
BR	0.960	0.0322	0.7160
BG	0.952	0.0321	0.7567

Table 3-2 95% Confidence Intervals of p

(R=1000, B=2000, n=1000, m=100 ; $\exp(\alpha_1)$=2.0, λ_1=0.5)

Method	Cov. Prob.	Length	Shape
T=1, Pr[D|T]=0.0952			
NR	0.950	0.0828	1.4061
NG	0.950	0.0813	1.4020
PR	0.923	0.0845	1.4280
PG	0.926	0.0818	1.3663
BR	0.961	0.0817	0.8939
BG	0.959	0.0800	0.8449
T=2, Pr[D|T]=0.1739			
NR	0.956	0.2407	1.5871
NG	0.956	0.2355	1.5845
PR	0.938	0.2546	1.7572
PG	0.942	0.2470	1.7196
BR	0.943	0.2356	1.0770
BG	0.942	0.2294	1.0500

Examination on confidence interval of *p*

(1) Cov. Prob. for Percentile :

does not keep the nominal level,

especially in case of $e^{\alpha_1} = 3.0$.

Cov. Prob. for BCa :

quite satisfactory for almost all cases.

(2) Length of C.I. is minimum when T=0.

It becomes large rapidly with increase of T.

(3) Length for NG is always shorter than that for NR.

References

[1] Carroll, R.J., Ruppert, D., Stefanski, L.A.: Measurement Error in Nonlinear Models, Chapman & Hall, New York (1995).

[2] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap, Chapman & Hall, New York (1993).

[3] Rosner, B, Willett, W.C., Spiegelman D.: Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, Statistics in Medicine, **8**, 1051--1069 (1989).