Bootstrap estimation of disease incidence proportion with measurement errors

Masataka Taguri[1] and Hisayuki Tsukuma[2]

[1] Graduate School of Medicine, University of Tokyo, taguri@epistat.m.u-tokyo.ac.jp

[2] Faculty of Medicine, Toho University, tsukuma@med.toho-u.ac.jp

Summary. This paper first treats the point and interval estimation of α_1 which specifies the odds ratio e^{α_1} in a disease incidence proportion with measurement errors. Some estimators are proposed, then the accuracy and the stability are investigated together with the well-known estimator proposed by Rosner et al. [RWS89]. Next the confidence interval for the disease incidence proportion is found by the normal approximation and the bootstrap methods. Numerical examinations are also carried out and some findings are given.

Key words: bootstrap interval estimation; inverse regression; odds ratio.

1 Introduction

It has long been appreciated that error in the measurement of individual exposure that is random (nondifferential) with respect to disease status will tend to bias estimates of relative risk towards the null value in most commonly considered instances. This causes mainly by random error of exposure measurement.

Researchers have attempted to correction and get unbiased estimates of disease risk. Initially, it was assumed that measurement errors were associated with only within person random variation. But in nutritional studies, for example, the use of the FFQ (Food Frequency Questionnaires) is likely to result in systematic error.

To obtain a corrected risk estimate, Rosner et al. [RWS89] introduced the linear Regression Calibration (RC) method, which estimates the attenuation coefficient as the slope of true on observed exposure. Because this approach allows for systematic bias as well as within person random variation in the FFQ, it has gained recognition as the best currently available approach for correcting risk estimates for dietary measurement error.

In this paper, we first investigate the accuracy and the stability of the risk estimates. We propose some risk estimators and compare them to the Rosner's estimator numerically. We also investigate the interval estimation based upon these estimators. Another interest in this study is the interval estimation of the incidence proportion, and we apply the normal approximation and the bootstrap methods to find the confidence intervals of the proportion via the estimators stated above.

In section 2, a brief review of the RC method is given. The properties of the risk estimators and the computational methods of confidence intervals are discussed in section 3. The results of simulation studies are presented in section 4, and some discussion is also made in the section.

2 Brief review of Regression Calibration (RC) method

Suppose the model relating a single-dimensional true exposure T and the probability of disease D is of logistic form;

$$\mathrm{logit}[\Pr(D|T)] = \alpha_0 + \alpha_1 T. \tag{1}$$

Further, assume that a linear relationship exists between true exposure T and observed exposure Q, that is, the following model is satisfied;

$$T = \lambda_0 + \lambda_1 Q + \epsilon, \quad \epsilon \sim N(0,\ \sigma^2). \tag{2}$$

We assumed throughout that a measurement error is nondifferential with respect to disease, i.e, $\Pr(D|T,\, Q) = \Pr(D|T)$. Ignoring measurement error, the logistic regression model for D on Q is the form $\mathrm{logit}[\Pr(D|Q)] = \gamma_0 + \gamma_1 Q$.

Now our objective is to determine α_1 from a data set (a) a main study population with observed Q exposure levels and disease statuses, and (b) a validation study population, distinct from (a), where both true T and observed Q exposure levels are available.

Using the RC method, we can consistently estimate α_1, true log odds ratio, as follows.

(i) Estimate γ_1 (denote $\hat{\gamma}_1$) by a maximum likelihood method from the main study data.
(ii) Estimate λ_1 (denote $\hat{\lambda}_1$) by ordinary least squares from the validation study data.
(iii) Estimate α_1 by $\hat{\alpha}_1 = \hat{\gamma}_1/\hat{\lambda}_1$ (see [RWS89]).

It has been shown through theory as well as through a detailed simulation study that when the disease is rare, the relative risk is not large, this estimator will remove most of the bias due to measurement error in Q. Applying the delta method, the asymptotic variance of $\hat{\alpha}_1 = \hat{\gamma}_1/\hat{\lambda}_1$ can be obtained as follows after some complicated manipulation;

$$Var(\hat{\alpha}_1) \approx \mathrm{Var}(\hat{\gamma}_1)/\lambda_1^2 + \gamma_1^2 \mathrm{Var}(\hat{\lambda}_1)/\lambda_1^4. \tag{3}$$

3 Estimation of α_1 and p

3.1 Stability of the point estimator $\hat{\alpha}_1$

The RC estimator of α_1 is given by $\hat{\alpha}_1 = \hat{\gamma}_1/\hat{\lambda}_1$. For a ratio estimator like $\hat{\alpha}_1$, if the coefficient of variations of the numerator and denominator are larger than 0.1 or if the sample size is small, for example less than 30, the variance of the ratio estimator tends to give a too low value and the positive skewness in the distribution may become noticeable (see [C77]). Moreover even with bivariate normality of the numerator and denominator, the confidence intervals for the ratio estimator derived from Fieller's limits have been criticized as not conservative enough (see also [C77]). These facts suggest that the RC estimator of α_1 might be unstable.

We therefore investigate more stable estimators of α_1. Let us regard (2) as the regression of T_i on Q_i, and consider the inverse regression model $Q_i = \beta_0 + \beta_1 T_i + \epsilon_i$. The ordinary least squares estimator of β_1 is then given by $\hat{\beta}_1 = \hat{\lambda}_1/\{(S_e/S_{QQ}) + \hat{\lambda}_1^2\}$, where $\hat{\lambda}_1$ is the ordinary least squares estimator calculated from (2). S_e and S_{QQ} are given by $S_e = \sum_{i=1}^m (T_i - \hat{\lambda}_0 - \hat{\lambda}_1 Q_i)^2$ and $S_{QQ} = \sum_{i=1}^m (Q_i - \bar{Q})^2$ respectively, where $\bar{Q}$ is the sample mean of Q_i's. Note that we can show the distribution of this estimator $\hat{\beta}_1$ has moments. Therefore if we use $\hat{\beta}_1$ instead of $1/\hat{\lambda}_1$, it may be possible to estimate α_1 with more stability. Extending this idea, consider the following ridge type estimator; $\hat{\beta}_{1k} = \hat{\lambda}_1/\{k\hat{\sigma}^2/S_{QQ} + \hat{\lambda}_1^2\}$, where k is a constant and $k/S_{QQ} = O(1)$ and $\hat{\sigma}^2 = S_e/(m-2)$. Based upon this idea we propose the following generalized estimator of α_1, which is obtained by similar way to the RC method;

$$\hat{\alpha}_{1k} = \hat{\gamma}_1 \hat{\beta}_{1k} = \hat{\gamma}_1 \times \hat{\lambda}_1/\{k\hat{\sigma}^2/S_{QQ} + \hat{\lambda}_1^2\}. \tag{4}$$

Note that $\hat{\alpha}_{1k}$ is the product of $\hat{\gamma}_1$ and $\hat{\beta}_{1k}$. They are mutually independent and both of them have moments, so $\hat{\alpha}_{1k}$ has moments for any finite sample size. If $k = 0$ this is nothing but the RC estimator, since $\hat{\beta}_{10} = 1/\hat{\lambda}_1$. In the following section 4, we make a comparison between $\hat{\alpha}_1$ and $\hat{\alpha}_{1k}$ numerically.

3.2 Interval estimation of α_1

The parameter α_1 in (1) means the increment of logit(Pr($D|T$)) when exposure increases by one unit, and e^{α_1} is the odds ratio for one unit increase of exposure. Hence we are interested in the inference of α_1. Rosner et al. [RWS89] examined the confidence intervals of α_1 by the normal approximation method. However, as mentioned in section 3.1, we should pay attention to estimating the asymptotic variance of $\hat{\alpha}_1$ based upon the delta method (see (3)), and to making the confidence intervals by normal approximation.

On the other hand $\hat{\alpha}_{1k}$ may be stable, hence it may be possible to find a good confidence interval based upon the normal approximation method.

We next consider to find the confidence interval of α_1 via the bootstrap methods. First we calculate $\hat{\alpha}_1^*$'s from bootstrap samples and then based upon the bootstrap distribution of $\hat{\alpha}_1^*$, we make the confidence interval of α_1. In case of the RC method, $\hat{\alpha}_1 = \hat{\gamma}_1/\hat{\lambda}_1$, where $\hat{\gamma}_1$ is calculated from a main sample and $\hat{\lambda}_1$ is from a validation sample. In this study, we first calculate $\hat{\gamma}_1^{*b}$ ($b = 1, 2, \ldots, B$) by resamples from the main sample. We also calculate $\hat{\lambda}_1^{*b}$ ($b = 1, 2, \ldots, B$) by resamples from the validation sample. Then they produce $\hat{\alpha}_1^{*b} = \hat{\gamma}_1^{*b}/\hat{\lambda}_1^{*b}$ for $b = 1, 2, \ldots, B$. In the numerical experiment of the next section 4, we consider the interval estimation based upon $\hat{\alpha}_{1k}$ in addition to $\hat{\alpha}_1$, where $\hat{\alpha}_{1k}$ is the proposed estimator (4). To find bootstrap confidence intervals, we consider the percentile method and the BC$_a$ method. For detailed algorithm of these methods, see Efron and Tibshirani [ET93].

3.3 Interval estimation of p

In preceding studies, researchers' main interest is to examine the interval estimation of α_1. However, we also have an interest in the disease incidence proportion $p =$

$\Pr(D|T)$, and the examination on the interval estimation of p must be important. In this subsection we discuss this problem.

Let us now consider to find the confidence interval of p by normal approximation and by bootstrap. From (1), a naive estimator of p is given by $\hat{p} = e^{\hat{\alpha}_0+\hat{\alpha}_1 T}/\{1 + e^{\hat{\alpha}_0+\hat{\alpha}_1 T}\}$, where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are estimated by the RC method or the generalized inverse regression method (4). Note that p should exist between 0 and 1, so we first find the confidence interval of $\text{logit}(p)$ and then obtain that of p by inverting $\text{logit}(p)$.

4 Numerical examination

4.1 Set-up of simulation study

As for the estimation of α_1, Rosner et al. [RWS89] carried out a simulation study. So we will carry out our simulation under similar condition to Rosner's for comparison. The procedure is as follows.

(i) From a standard normal distribution $N(0,1)$, generate T_i for $i = 1, \ldots, m$, where m is the size of a validation sample.

(ii) Assume the model $Q_i = T_i + e_i, \quad e_i \sim N(0, (1-\lambda_1)/\lambda_1)$. First generate e_i for $i = 1, 2, \ldots, m$. Then by combining these e_i's with T_i's generated in the step (i), make a validation sample $(T_i,\ Q_i) \quad (i = 1, 2, \ldots, m)$. For the value of λ_1, the computation is done in case of $\lambda_1 = 0.3,\ 0.5,\ 0.7$, because these values seem to often appear in practical situations.

(iii) A main sample $(Y_i,\ Q_i)$ with size n is made by the following procedure: First similar to the steps (i) and (ii), generate $(T_i,\ Q_i) \quad (i = 1, 2, \ldots, n)$. For the value of α_1, we consider the cases $e^{\alpha_1} = 1.5,\ 2.0,\ 3.0,\ 4.0,\ 5.0$, where these values may often appear in practical situations. The value of α_0 is considered to be the value of $\text{logit}(p)$ for $T = 0$. So we fix the value so as to satisfy $p = 0.05$, that is, the relation $e^{\alpha_0}/(1+e^{\alpha_0}) = 0.05$ is satisfied. Using the values of α_0 and α_1, calculate the value of p given T_i. Then generate Y_i from the Bernoulli distribution with the parameter p computed above. By combining these Y_i's with Q_i's which are the paired values of T_i's, make a main sample $(Y_i,\ Q_i) \quad (i = 1, 2, \ldots, n)$.

(iv) Set $T = -1,\ 0,\ 1,\ 2$ for which the disease incidence proportion should be obtained. Then compute the confidence intervals with 95 percent confidence coefficient by normal approximation, percentile or BC$_a$ method given in section 3.

In the following simulation study, we set $m = 100$, $n = 1,000$, and the number of simulations $R = 1,000$. The computation is carried out for all combination of the values λ_1 and α_1.

Remark. The model considered in the above step (ii) seems to be different from the model (2) given in section 2. However it is shown that the model is the special case of being $\lambda_0 = 0$ and $\sigma^2 = 1 - \lambda_1$ in (2).

4.2 Examination on stability of the estimators for α_1

In this study we consider two types of estimator for α_1; $\hat{\alpha}_1^R$ is the RC estimator given in section 2, and $\hat{\alpha}_1^G$ is the generalized estimator given by (4) where $k = 1$. For the values of α_1 and λ_1, we consider $e^{\alpha_1} = 1.5, 2.0, 3.0, 4.0, 5.0$, and $\lambda_1 = 0.3, 0.5, 0.7$, respectively. The number of bootstrap replication is $B = 2,000$.

Table 1 shows a part of the computational results. That is, we compute the mean, standard deviation (SD), skewness, kurtosis, minimum and maximum values, and the range over 1,000 simulations for two kinds of the estimates. From this table, we can get the following findings.

(1) $\hat{\alpha}_1^G$ underestimates the true value except for the case $e^{\alpha_1} = 1.5, \lambda_1 = 0.7$. $\hat{\alpha}_1^G$ and $\hat{\alpha}_1^R$ are monotone increasing with respect to λ_1 except $\hat{\alpha}_1^R$ for $e^{\alpha_1} = 1.5$ and 2.0. As for the comparison of $\hat{\alpha}_1^G$ and $\hat{\alpha}_1^R$, $\hat{\alpha}_1^G$ is smaller than $\hat{\alpha}_1^R$ for all cases. The biases of $\hat{\alpha}_1^G$ and $\hat{\alpha}_1^R$ are not so different on the whole.

(2) The values of SD, skewness, kurtosis and range for $\hat{\alpha}_1^G$ are almost all smaller than those for $\hat{\alpha}_1^R$. The differences of skewness and kurtosis are remarkable for small value of λ_1. In case of $\lambda_1 = 0.3$, they range from 6.0% to 10.9% for skewness, and from 0.9% to 4.6% for kurtosis. So the tail of $\hat{\alpha}_1^R$'s distribution is heavier than that of $\hat{\alpha}_1^G$'s, especially when λ_1 is small.

4.3 Examination on confidence interval of α_1

Table 2 shows a part of the computational results for the interval estimation of α_1. That is, we compute the coverage probability, length, shape over 1,000 simulations for two kinds of estimates, where confidence coefficient is 95 percent. The shape is defined by $(c_U - \hat{\alpha}_1)/(\hat{\alpha}_1 - c_L)$, where c_L and c_U are the lower and the upper limits of the confidence interval, respectively. The first column of the table indicates estimation methods; N$*$, P$*$, and B$*$ are corresponding to Normal approximation, Percentile, and BC$_a$ methods, respectively. $*$R and $*$G are corresponding to $\hat{\alpha}_1^R$ and $\hat{\alpha}_1^G$, respectively. From this table, we can get the following findings.

(1) The coverage probability of NG is always smaller than that of NR. The coverage probability of Normal approximation method is monotone decreasing with respect to true odds ratio.

(2) The length of the confidence interval based upon $\hat{\alpha}_1^G$ is shorter than that based upon $\hat{\alpha}_1^R$ for all cases. This suggests the stable property of $\hat{\alpha}_1^G$.

(3) From the viewpoint of coverage probability, Normal approximation method is slightly better than Percentile method for small values of α_1 ($e^{\alpha_1} = 1.5, 2.0$). However Percentile is much better than Normal approximation for large values of α_1 ($e^{\alpha_1} = 3.0, 4.0, 5.0$); especially the coverage probabilities of Normal approximation are less than 80% for the case of $e^{\alpha_1} = 5.0$ and $\lambda_1 = 0.3, 0.5$.

(4) BC$_a$ method is worse than other methods. This reason may be that the bias and/or skewness of the estimates' distribution are too much adjusted by BC$_a$ method.

4.4 Examination on confidence interval of p

Table 3 shows a part of the computational results for the 95 percent confidence intervals of $p = \Pr(D|T)$. That is, we compute the coverage probability, the length and shape of confidence intervals. From this table, we can get the following findings.

(1) The coverage probability for Normal approximation and Percentile method does not keep the nominal level in some cases, especially for large odds ratio. On the other hand, the coverage probability for BC$_a$ method is quite satisfactory except for $T = 0$.

(2) The length of confidence interval tends to become large with the increase of T ; especially for $T = 1, 2$.

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

References

[C77] Cochran, W.G.: Sampling Techniques (third edition). John Wiley & Sons, Inc., New York (1977)

[ET93] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

[RWS89] 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)

Table 1. Estimates of α_1 ($R = 1,000, B = 2,000, n = 1,000, m = 100$).

Estimator	Mean	SD	Skewness	Kurtosis	Min.	Max.	Range
$\alpha_1 = 0.4054\,(e^{\alpha_1} = 1.5),\ \lambda_1 = 0.5$							
$\hat{\alpha}_1^R$	0.4097	0.2140	0.2274	3.2981	−0.3239	1.2488	1.5727
$\hat{\alpha}_1^G$	0.4052	0.2115	0.2243	3.2971	−0.3187	1.2377	1.5564
$\alpha_1 = 1.0986\,(e^{\alpha_1} = 3.0),\ \lambda_1 = 0.5$							
$\hat{\alpha}_1^R$	1.0524	0.2166	0.4132	3.2349	0.5157	1.9238	1.4081
$\hat{\alpha}_1^G$	1.0409	0.2130	0.4021	3.2127	0.5122	1.8921	1.3798
$\alpha_1 = 1.6094\,(e^{\alpha_1} = 5.0),\ \lambda_1 = 0.5$							
$\hat{\alpha}_1^R$	1.4132	0.2274	0.5507	3.5719	0.7604	2.4117	1.6513
$\hat{\alpha}_1^G$	1.3978	0.2229	0.5364	3.5317	0.7546	2.3725	1.6179

Table 2. 95 percent Confidence intervals of α_1 by the normal approximation and the bootstrap methods ($R = 1,000, B = 2,000, n = 1,000, m = 100$).

Method	Coverage probability	Length	Shape
$\alpha_1 = 0.4054\,(e^{\alpha_1} = 1.5)$, $\lambda_1 = 0.5$			
NR	0.956	0.8212	1.0000
NG	0.954	0.8116	1.0000
PR	0.942	0.8495	1.1323
PG	0.945	0.8377	1.1265
BR	0.958	0.8846	0.6939
BG	0.956	0.8729	0.6920
$\alpha_1 = 1.0986\,(e^{\alpha_1} = 3.0)$, $\lambda_1 = 0.5$			
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
$\alpha_1 = 1.6049\,(e^{\alpha_1} = 5.0)$, $\lambda_1 = 0.5$			
NR	0.804	0.8889	1.0000
NG	0.779	0.8727	1.0000
PR	0.876	0.9292	1.4583
PG	0.863	0.9062	1.4384
BR	0.746	0.9070	0.8034
BG	0.721	0.8882	0.7960

Table 3. 95 percent Confidence intervals of $\Pr(D|T)$ by the normal approximation and the bootstrap methods ($R = 1,000, B = 2,000, n = 1,000, m = 100, \lambda_1 = 0.5$).

	$e^{\alpha_1} = 2.0$			$e^{\alpha_1} = 3.0$		
Method	C.P.†	Length	Shape	C.P.†	Length	Shape
$T = -1$	$p = 0.0256^\ddagger$			$p = 0.0172^\ddagger$		
NR	0.917	0.0342	1.7567	0.813	0.0300	1.8303
NG	0.919	0.0341	1.7481	0.799	0.0300	1.8190
PR	0.933	0.0318	1.2882	0.862	0.0272	1.3086
PG	0.933	0.0321	1.3169	0.851	0.0278	1.3587
BR	0.954	0.0314	0.8901	0.935	0.0268	0.9270
BG	0.954	0.0315	0.9193	0.932	0.0271	0.9747
$T = 0$	$p = 0.05^\ddagger$			$p = 0.05^\ddagger$		
NR	0.905	0.0333	1.3310	0.711	0.0415	1.3574
NG	0.905	0.0332	1.3301	0.710	0.0413	1.3557
PR	0.919	0.0313	1.0245	0.725	0.0371	1.0576
PG	0.927	0.0314	1.0585	0.726	0.0378	1.1185
BR	0.960	0.0322	0.7160	0.870	0.0382	0.7120
BG	0.952	0.0321	0.7567	0.850	0.0383	0.7774
$T = 1$	$p = 0.0952^\ddagger$			$p = 0.1364^\ddagger$		
NR	0.950	0.0828	1.4061	0.914	0.1224	1.3409
NG	0.950	0.0813	1.4020	0.919	0.1197	1.3378
PR	0.923	0.0845	1.4280	0.847	0.1243	1.5806
PG	0.926	0.0818	1.3663	0.863	0.1201	1.4922
BR	0.961	0.0817	0.8939	0.945	0.1175	0.9673
BG	0.959	0.0800	0.8449	0.954	0.1157	0.8877
$T = 2$	$p = 0.1739^\ddagger$			$p = 0.3214^\ddagger$		
NR	0.956	0.2407	1.5871	0.969	0.3384	1.2272
NG	0.956	0.2355	1.5845	0.969	0.3318	1.2334
PR	0.938	0.2546	1.7572	0.940	0.3527	1.5395
PG	0.942	0.2470	1.7196	0.940	0.3444	1.5075
BR	0.943	0.2356	1.0770	0.964	0.3393	0.9383
BG	0.942	0.2294	1.0500	0.968	0.3333	0.9075

† C.P. : Coverage Probability

‡ p : True value of $\Pr(D|T)$