

Comparison of the Breslow-Holubkov ML estimator to other odds ratio estimators in two-phase and counter-matched studies

Yu-Fen Li ¹ and Bryan Langholz ¹

Abbreviation: NONE

¹ Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California

Correspondence to:
Yu-Fen Li, MS MPH
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
1540 Alcazar Street, Suite 236
Los Angeles, CA 90033
Email: yufenli@usc.edu
Phone: (323) 442-2772
Fax: (323) 442-3272

RUNNING HEAD: Breslow-Holubkov ML estimator vs other estimators
KEY WORDS: frequency matching, counter matching, and two-phase sampling

Abstract

Langholz and Goldstein compared the relative efficiencies of frequency matching, counter matching, and balanced two phase designs with full cohort using conditional and unconditional logistic analyses. There is some efficiency advantage of the conditional over the marginal (unconditional) likelihood for case-control studies with complex sampling. Breslow and Holubkov considered an maximum likelihood approach for getting estimates. The aim of this work is to expand the finding of Langholz and Goldstein including the maximum likelihood method proposed by Breslow and Holubkov.

A simulation study was performed on frequency matching, counter matching, and two-phase data. We found the variance of $\hat{\beta}$ estimated by frequency matching was larger than the other designs. Two-phase and counter matching designs are about same in parameter estimation and its variance. A larger number of iterations for the marginal likelihood analysis does not dramatically improve the estimation.

1 Introduction

Case-control studies are the most commonly used study design in disease epidemiology to assess the association of exposure with disease. Typically we select most of cases and just a small fraction of controls. There are several different ways to sample controls. Langholz and Goldstein (2001) compared the relative efficiencies of frequency matching, counter matching, and balanced two-phase designs with full cohort using conditional and unconditional logistic analyses. For case-control studies with complex sampling, the efficiency of conditional likelihood method is better than that of marginal (unconditional) likelihood method.

In this report, focus was on the maximum likelihood (ML) method for two-phase studies proposed by Breslow and Holubkov (1997). They demonstrated an efficiency advantage for the ML estimator. Augmenting the simulation studies presented in Langholz and Goldstein (2001), we compared it with maximum marginal and, where appropriate, conditional likelihoods.

2 Methods

In this section we briefly introduce frequency matching, counter matching and two-phase sampling, as well as the ML method evaluated in this report.

2.1 Frequency Matching

Frequency matching is a commonly used design, in which the number of controls is proportional to the number of cases. Suppose we sample md from the $n - d$ controls in one case-control set (e.g. one sampling stratum). For 1:m frequency matching, the probability that \mathbf{r} is the sampled case-control set given that the case set is \mathbf{d} would be

$$\pi(\mathbf{r}|\mathbf{d}) = \binom{n-d}{md}^{-1}$$

for each set \mathbf{r} of size $(m+1) \times d$ containing \mathbf{d} .

2.2 Counter Matching

It is assumed that the sampling strata variable $S \in \{1, \dots, J\}$ is known for all subjects in the study base. In counter-matching, the marginal total in

the sampling stratum j is fixed to a value proportional to the number of cases, $m_j d$. Then, $m_j d - d_j$ controls are randomly sampled without replacement from the $n_j - d_j$ total controls in stratum j . Counter-matching control selection is characterized by

$$\pi(\mathbf{r}|\mathbf{d}) = \left[\prod_{j=1}^J \binom{n_j - d_j}{m_j d - d_j} \right]^{-1}$$

for \mathbf{r} with $|\mathbf{r}_j| = m_j d$, $j = 1, \dots, J$.

2.3 Two-Phase Sampling

Let Y denote a random response variable with values 0 and 1 for cases and controls, respectively. Let X be a p -vector of explanatory variables with values $X = x_k$ for $k = 1, \dots, K$. Population is stratified into J strata, with the j^{th} stratum indexed by $S = j$ for $j = 1, \dots, J$. A linear logistic model is assumed,

$$\text{pr}(Y = 1|S = j, X = x_k) = \frac{\exp(x_k^T \beta)}{1 + \exp(x_k^T \beta)}$$

where β denotes a p -vector of regression coefficients which are logarithm of odds ratios (OR).

Suppose that N_0 controls ($i = 0$) and N_1 cases ($i = 1$) are randomly drawn from the infinite subpopulations. In the second phase, n_{ij} random subsamples are drawn from among N_{ij} within each of the $2 \times J$ strata, for $i = 0, 1$ and $j = 1, \dots, J$. Then, explanatory variables are measured and the numbers n_{ijk} with $X = x_k$ are determined.

Breslow and Holubkov (1997) derived an ML method for estimating OR, i.e. $\exp(\beta)$. In stratum j , the linear logistic regression is with predictor $x_k^T \beta + \xi_j$, instead of $x_k^T \beta$. ξ_j is defined as

$$\xi_j = \log\left(\frac{N_1}{N_0}\right) + \log\left(\frac{F_j}{n_{+j} - F_j}\right) + \log\left(\frac{N_{0j} + n_{1j} - F_j}{N_{1j} - n_{1j} + F_j}\right)$$

where F_j is the sum of the fitted values in stratum j from this logistic fit to the second phase data, and $+$ denotes summation over the corresponding index. Set initial values of $F_j = n_{1j}$. The ML estimator may be obtained from repeated logistic regression fits to the phase two data with recalculation of the ξ_j .

2.4 Simulations

In a simulation, cohort consisted of 10 study bases with 300 subjects each. The overall probability of disease was about 10%, an average of 10 cases per study base. Three different probabilities of exposure were considered: 50%, 20%, and 5%. From each simulated cohort, a 1:1 frequency matching, a 1:1 counter matching, and a 1:1 two-phase sample were randomly drawn. For counter matching, sampling stratum variable had 90% sensitivity and specificity for the exposure variable. For the two-phase sampling, all cases are selected at the second phase, i.e. $n_{1j} = N_{1j}$.

For each frequency- and counter-matched data set, the odds ratio was estimated using the appropriate conditional and unconditional logistic likelihoods. For the marginal likelihoods, a baseline odds parameter was estimated for each study base. We also tried a larger number of iterations (set as 8) for the marginal likelihood approach on counter matching and two-phase designs. The conditional likelihood was fitted using Procedure PHREG of SAS v8.1 by treating each set as an “individual” with covariate value equal to the sum of the set covariates. The “case” in this pseudo case-control data corresponds to the case set and each “control” corresponds to a non-case set. Details were described in Section 6.1 of Langholz and Goldstein (2001). Procedure LOGISTIC of SAS v8.1 was used for the analyses of full cohort and two-phase data.

3 Results

Simulation results were based on 1,000 trials and are shown in Table 1, 2 and 3.

When the exposure is not rare (i.e. Table 1 and 2), there is no evidence of bias in β estimates from the conditional logistic likelihood method. The marginal likelihood estimates with one iteration for frequency matching at OR = 4 is slightly biased. The variance of $\hat{\beta}$ from frequency matching was larger than the other designs. For frequency matching, the variance of $\hat{\beta}$ from the marginal likelihood is well estimated by the inverse information (I^{-1}) but not from counter-matched and two-phase data. Two-phase and counter matching designs are about same in parameter estimation and its variance. A larger number of iterations for the marginal likelihood analysis does not dramatically improve the estimation.

When the exposure is rare (Table 3), the variance of $\hat{\beta}$ estimated by the frequency matching method is also larger than the other designs. Moreover, I^{-1} from the marginal likelihood does not well estimate the variance of $\hat{\beta}$ for counter matching and two-phase models, as well as frequency matching which has a pretty good variance estimation when the exposure is not rare.

4 Discussion

The relative efficiencies of frequency matching, counter matching, and balanced two phase designs with full cohort using conditional and unconditional logistic analyses were compared by a simulation study. We could reproduce the results of Langholz and Goldstein. There is some efficiency advantage of the conditional over the marginal likelihood for case-control studies with complex sampling. There is no evidence of bias in β estimates from the conditional logistic likelihood method. The variance estimated by the counter-matched and two-phase data is smaller than that by the frequency-matched data. But we do not see the efficiency advantage of the ML method for the two-phase design. It might result from the difference in sampling fractions of cases and controls. First, we did not sample but take all cases at the second phase. Second, their study base was big and they sampled relatively small amount of cases (about 12%) and controls (about 0.4%). Therefore, our next step is to extend our simulations to case sampling at the second phase.

References

- [1] B. Langholz and L. Goldstein. Conditional logistic analysis of case-control studies with complex sampling. *Biostatistics*, 2:63-84, 2001.
- [2] N.E. Breslow and R. Holubkov. Weighted likelihood, pseudo-likelihood, and maximum likelihood methods for logistic regression analysis of two-phase data. *Statistics in Medicine*, 16:103-116, 1997.

Table 1: Results of simulation studies comparing cohort, frequency matching, counter matching, and two-phase designs using either conditional or marginal (unconditional) likelihood method. Cohorts consisted of 10 study bases with 300 subjects each. The probability of exposure was 50%, and the overall probability of disease was about 10%. Based on 1,000 trials.

	OR = 1 ($\beta = 0$)			OR = 4 ($\beta = 1.39$)		
	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}
Cohort	-0.00	0.015	0.015	1.39	0.024	0.021
Frequency matching						
Conditional	-0.00	0.027	0.027	1.38	0.035	0.033
Marginal (iter=1)	-0.00	0.028	0.027	1.41	0.036	0.033
Counter matching						
Conditional	-0.00	0.017	0.018	1.39	0.027	0.025
Marginal (iter=1)	-0.00	0.018	0.027	1.38	0.030	0.036
Marginal (iter=8)	-0.00	0.018	0.027	1.40	0.028	0.033
Two-phase						
Marginal (iter=1)	-0.00	0.019	0.027	1.39	0.027	0.036
Marginal (iter=8)	-0.00	0.018	0.027	1.40	0.027	0.033

Table 2: Results of simulation studies comparing cohort, frequency matching, counter matching, and two-phase designs using either conditional or marginal (unconditional) likelihood method. Cohorts consisted of 10 study bases with 300 subjects each. The probability of exposure was 20%, and the overall probability of disease was about 10%. Based on 1,000 trials.

	OR = 1 ($\beta = 0$)			OR = 4 ($\beta = 1.39$)		
	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}
Cohort	0.00	0.023	0.024	1.40	0.016	0.016
Frequency matching						
Conditional	0.01	0.039	0.042	1.40	0.036	0.037
Marginal (iter=1)	0.00	0.041	0.043	1.42	0.037	0.037
Counter matching						
Conditional	0.00	0.026	0.028	1.39	0.020	0.020
Marginal (iter=1)	0.03	0.027	0.039	1.40	0.022	0.029
Marginal (iter=8)	0.00	0.027	0.038	1.40	0.019	0.029
Two-phase						
Marginal (iter=1)	0.00	0.029	0.037	1.40	0.022	0.029
Marginal (iter=8)	0.00	0.028	0.037	1.40	0.020	0.029

Table 3: Results of simulation studies comparing cohort, frequency matching, counter matching, and two-phase designs using either conditional or marginal (unconditional) likelihood method. Cohorts consisted of 10 study bases with 300 subjects each. The probability of exposure was 5%, and the overall probability of disease was about 10%. Based on 1,000 trials.

	OR = 1 ($\beta = 0$)			OR = 4 ($\beta = 1.39$)		
	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}	$\hat{\beta}$	$Var(\hat{\beta})$	I^{-1}
Cohort	-0.03	0.088	0.083	1.38	0.037	0.038
Frequency matching						
Conditional	0.00	0.159	0.150	1.42	0.134	0.125
Marginal (iter=1)	0.00	0.164	0.153	1.44	0.137	0.127
Counter matching						
Conditional	-0.03	0.099	0.094	1.38	0.044	0.046
Marginal (iter=1)	0.00	0.105	0.110	1.41	0.048	0.058
Marginal (iter=8)	-0.03	0.099	0.109	1.38	0.045	0.057
Two-phase						
Marginal (iter=1)	-0.03	0.103	0.105	1.39	0.052	0.061
Marginal (iter=8)	-0.02	0.103	0.106	1.39	0.052	0.061