Covariate Adjustment May Not Be Better: Thresholds of Relative Risk when Controlling for Confounders

Wenjun Li
Biostatistics Research Group, Division of Preventive and Behavioral Medicine, University of Massachusetts Medical School, Worcester, MA

Edward J. Stanek III
Department of Public Health, University of Massachusetts, Amherst, MA

Co-authors ??

Jenny Kelsey ??

Liz Bertone ??

Table of Contents

Abstract....................................................................................................................................2
Introduction ..........................................................................................................................3
Background .........................................................................................................................4
  Crude rates .....................................................................................................................6
  Covariate-adjusted rates ...............................................................................................6
Simulation methods ...........................................................................................................9
Results ............................................................................................................................10
Discussion .......................................................................................................................13
Reference ........................................................................................................................16
Abstract

In theory, covariate-adjusted rates have smaller variance with known variance components. In practice, variance components needed to achieve the adjustment are unknown and their sample estimators are used instead. We develop empirical guidelines to determine when adjusted rates have smaller variance than crude rates, illustrating the results in an example of adjusting for gender when estimating cigarette smoking prevalence. We show that in many practical settings, adjustment for gender when estimating prevalence rate should not be conducted. The variance reductions due to covariate adjustment depend on sample sizes, gender ratios and male-to-female prevalence rate ratios (RRs). In populations with balanced gender ratios and prevalence rates of 35%, the adjusted rates had smaller variances when RRs were above 1.6, 1.5, 1.4, 1.3 and 1.2 for sample sizes of 25, 50, 100, 150 and 200, respectively. In populations with unbalanced gender ratios or lower prevalence rates, the RR thresholds were higher. In sum, adjusted rates should not be used in all settings, and in particular, not when both RRs and sample sizes are small and prevalence rates are low. When reporting smoking rates at the municipal level based on Massachusetts BRFSS data, gender adjustment is not recommended when sample size is less than 200.

Key Words: Epidemiologic method, survey sampling, finite population, rate adjustment, variance components
Introduction

Epidemiologic studies often involve random sampling of study subjects from a population defined in space and time. For example, the Behavioral Risk Factors Surveillance Systems (BRFSS) of the Centers for Disease Control and Prevention (CDC) conducts annual telephone surveys on adults living in households in the 50 states of the United States. BRFSS surveys cover many behavioral risk factors, such as cigarette smoking, sexual behaviors, and drunk driving. When the sample data are analyzed, adjustments are usually made to account for possible imbalance of covariates (such as gender and age) in the study samples, which is also commonly referred as “control for confounding”. In theory, estimators can be made more precise by adjusting for known auxiliary information such as gender proportion and average age of the study population (Särndal et al. 1992). Adjusted rate estimators have smaller variance when relevant variance components are known (Särndal et al. 1992). However, in practice, variance components are unknown and replaced by their sample estimator. The uncertainty introduced by using variance component estimates has implications for whether controlling for a confounder effectively reduces the size of the confidence interval for prevalence.

In many epidemiologic or public health studies relying on random samples, relevant variance components are unknown. To our knowledge, there is no established guideline with regard to when the covariate-adjustment should or should not be made in popular epidemiology textbooks, such as Rothman and Greenland (1998) and Fleiss, Levin and Paik (2003). We present such guidelines. We illustrate using a simple example that covariate adjustment may lead to rate estimators with larger variance and thus wider confidence intervals than un-adjusted estimators when variance components
are unknown, and suggest an empirical guideline on when such adjustment should not be used in epidemiologic reports.

**Background**

We use a simple hypothetical example to introduce the problem and define notation. Suppose a survey of cigarette smoking prevalence rates among adult residents living in a small town in Massachusetts is conducted by random digit dialing following the procedures similar to the BRFSS (CDC 2004). The town list of residents, updated annually by the town government as mandated by law, provides telephone numbers of adult residents. The total number of adults ($N$) and the proportion of male adult residents $\pi_x$ are known. For simplicity, we suppose the probability of each household telephone number being called is proportional to the number of adults in the household and it is allowed to call the same telephone number multiple times if multiple adults with same telephone number are selected in the sample. This sampling process is thus a simple random sampling (SRS) that all adults have equal probability of being interviewed. In total, a sample of $n$ subjects is selected. We represent the smoking status for the $i$-th sample subject with an indicator variable $Y_i$ (1 if a subject is a smoker, and zero otherwise), and the auxiliary variable $X_i$ is an indicator of male gender (1 if the subject is male, and zero otherwise). We assume that smoking status and gender are recorded for all the subjects in the sample. The goal of the survey is to estimate the prevalence rate of cigarette smoking $\pi_y$. The data from the sample and information on the number of male and female residents are summarized into the following 2 x 2 table (Table 1). We estimate smoking rates using the crude and the gender-adjusted rates, both of which are unbiased (Cochran 1977; Särndal et al. 1992).
Table 1 Number of subjects in a simple random sample / a hypothetic population
by gender and smoking status

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Total</th>
<th>Sample</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>$n_{11}$</th>
<th>$n_{10}$</th>
<th>$n_{1*}$</th>
<th>$N_{1*}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>$n_{21}$</td>
<td>$n_{20}$</td>
<td>$n_{2*}$</td>
<td>$N_{2*}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1*}$</td>
<td>$n_{0*}$</td>
<td>$n$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Notation:

- $n_{1*}$ = number of smokers in the sample
- $n_{11}$ = number of male smokers in the sample
- $n_{21}$ = number of female smokers in the sample
- $n_{0*}$ = number of nonsmokers in the sample
- $n_{10}$ = number of male nonsmokers in the sample
- $n_{20}$ = number of female nonsmokers in the sample
- $n = n_{1*} + n_{0*} = n_{1*} + n_{2*}$ = sample size
- $N_{1*}$ = number of males in the population
- $N_{2*}$ = number of females in the population
- $N$ = population size
- $\pi_x = N_{1*}/N$ proportion of males in the population
Crude rates

The crude rates are straightforward estimators of the prevalence rate, i.e., sample mean of the indicator variable for smoking status $Y$. The crude rate and its variance are given by $\hat{\pi}_y = \bar{Y} = n_1/n$ and $\text{var}(\hat{\pi}_y) = (1 - f) \sigma^2_y/n$, where $f = n/N$, and $\sigma^2_y = \pi_y (1 - \pi_y)$ is variance of $Y$.

Covariate-adjusted rates

The covariate-adjusted rate estimator corresponds to a simple weighting of the gender specific rates according to gender proportions of the population. With the data in Table 1, the gender-adjusted rate estimator is calculated as

$$\hat{\pi}_y = \frac{N_{1*}}{N} \left( \frac{n_{11}}{n_{*}} \right) + \frac{N_{2*}}{N} \left( \frac{n_{21}}{n_{*}} \right).$$

(1)

This estimator, routinely used in epidemiological and public health reports, is referred as the poststratified estimator in finite sampling literature and has been shown using both calibration (Deville and Särndal 1992) and model-assisted approaches (Särndal et al. 1992).

Although the adjustment is intuitive and seemingly straightforward, the underlying variability resulting from sampling both $y$ and $x$ results in variability that may overshadow the gains made by the adjustment. To examine this issue, it is valuable to consider more general statistical frameworks that give rise to this solution.

Various approaches have been proposed in finite sampling literature to obtain covariate-adjusted estimates of rates based on simple random samples. For example, methods of improving estimation with auxiliary information have been discussed in a model-based approach (Bolfarine and Zacks 1992; Valliant et al. 2000), model-assisted
(Cassel et al. 1977; Cochran 1977; Särndal et al. 1992) or calibration methods (Särndal and Wright 1984; Deville and Särndal 1992), and design-based prediction approach (Li and Stanek III 2006, under review). For the simple scenarios concerned in this paper, the above approaches result in very similar estimators, which vary slightly according to what variance components are used in calculation. For example, with design-based prediction approach by Stanek and Singer (2005) and Li and Stanek (2006, under review), the gender-adjusted estimator is derived as

\[
\hat{\pi}_y = \frac{1}{N} \left\{ \sum_{i=1}^{n_i} Y_i + \sum_{i=n+1}^{N} \left( Y - \frac{1}{1-f} \beta (\bar{X} - \pi_x) \right) \right\},
\]

and its variance

\[
\text{var} \left( \hat{\pi}_y \right) = (1 - \rho^2)(1 - f) \sigma_y^2 / n,
\]

where \( f = n/N \), \( \beta = \sigma_{xy} / \sigma_x^2 \), \( \rho = \sigma_{xy} / \sigma_x \sigma_y \) is the correlation coefficient of smoking status \( Y \) on gender \( X \), and \( \bar{Y} \) is the sample proportion of cigarette smokers in the sample and \( \bar{X} \) is the proportion of male subjects in the sample, respectively.

In practice, these quantities can be estimated directly from the sample. For example, with the data in Table 1, \( \bar{Y} = n_1/n \), \( \bar{X} = n_*/n \), and \( \sigma_x^2 = N_1 N_2/\left( N(N-1) \right) \). For \( \sigma_y^2 \) and \( \sigma_{xy} \), we apply “method of moments” and replace \( \sigma_y^2 \) and \( \sigma_{xy} \) with their sample estimates, \( \hat{\sigma}_y^2 = \frac{n_1 n_{*0}}{n(n-1)} \) and \( \hat{\sigma}_{xy} = \frac{n_1 n_{2*}}{n(n-1)} \left( \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2} \right) \), and obtain

\[
\hat{\beta}_{xy} = \left( \frac{N(N-1)}{n(n-1)} \right) \left( \frac{n_1 n_{2*}}{N_1 N_2} \right) \left( \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2} \right). \]
Alternatively, if $\sigma_y^2$, $\sigma_{xy}$ and $\sigma_x^2$ are all replaced with their sample estimates $\hat{\sigma}_y^2$, $\hat{\sigma}_{xy}$ and $\hat{\sigma}_x^2 = \frac{n_1 n_2}{n(n-1)}$, then

$$\hat{\beta}_{xy} = \frac{n_{11} - n_{21}}{n_1^* - n_2^*},$$

and consequently, we obtain the post-stratified estimator (see equation (1)).

The gender-adjusted rate estimator in equation (2) has interesting interpretations. The first expression is divided into the sum of random variables in the observed sample and a predictor of the random variables in the unobserved subjects of the population. Since the weighted total of the sample and the unobserved random variables equals the population mean, it is clear that for a realized sample, we simply substitute the values observed for the sample random variables in the estimator. The adjusted rates, in either the first or the second expressions, include the sample prevalence rate, $\bar{Y}$, and an adjustment term for gender. The adjustment ‘shrinks’ the observed difference in male proportions between the sample and the population. For example, if the value observed for the male proportion in the sample, $\bar{X}$, exceeds the population male proportion, $\pi_x = N_1/N$, and the cigarette smoking status and gender random variables are positively correlated, we may anticipate that the sample smoking prevalence, $\bar{Y}$, will also exceed the population smoking prevalence $\pi_y$. The adjusted rate compensates for the over estimation of $\pi_y$ by the sample random variables, and the adjustment is proportional to the difference in male proportions, $\bar{X} - \pi_x$, and weighted by the regression coefficient, $\beta$. As shown in formula (3), the adjustment results in variance
reduction over the simple sample proportion, with the percent reduction given by
\[(1 - \rho^2) \times 100\%\].

The adjusted estimator (2) is a function of the regression coefficient, which, in
turn, is a function of variance components in the population. In practice, although \(\sigma_x^2\)
may be known, \(\sigma_{xy}\) will need to be estimated. Särndal et al. (1992, p229) recommends
estimating both terms, a recommendation that we evaluate via simulation. The estimate
of \(\beta\) is given by \(b_1 = \hat{\sigma}_{xy} / \hat{\sigma}_x\) (see formula (5)), and the resulting adjusted rate estimator
is the well-known post-stratified estimator (1). Li and Stanek (2006, under review)
showed that use of known population variance of \(X\) to estimate \(\beta\) by \(b_2 = \hat{\sigma}_{xy} / \sigma_x\)
(formula (4)) generally result in smaller variance of the adjusted rates, especially when
sample sizes are relatively small. In this analysis, gender-adjusted rates were made by
using \(b_2\).

For the adjusted estimators, we examine how the uncertainty of estimating
variance components influence the precision of rate estimates. To do so, we conducted
a series of Monte Carlo simulations to compare the variances of the crude and the
adjusted rates.

**Simulation methods**

The simulation study generated a series of hypothetical populations of sizes 200,
400, 800, 1,600 and 3,200; each with male proportions ranging from 30% to 70% (\(\pi_x =
0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70\)), and with prevalence rates of
cigarette smokers ranging from 15% to 35% (\(\pi_y = 0.15, 0.20, 0.25, 0.30\) and 0.35). The
prevalence rate ratios (RR) of men to women ranged from 1.0 to 4.0 (RR = 1.0, 1.1, 1.2,
1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 2.0, 2.2, 2.4, 3.0, 3.5, 4.0). According to data from 2004 BRFSS data of Massachusetts (http://apps.nccd.cdc.gov/brfss/), the prevalence of current cigarette smokers is 18.4%; male-to-female RR is 1.13, RR of adults with low annual per ca income (< $15,000) to those with high annual per ca income (> $50,000) is 2.1; the RR of individuals with less than high school education to college graduates is 2.6; and the RR of those aged 18 to 25 to those aged 65 or older is 3.1. The ranges of gender ratios and smoking prevalence rates are similar to those in 1999-2003 BRFSS data of Massachusetts.

We evaluated the rate estimators by comparing the average variance over 10,000 independent simple random samples for each scenario. The sample sizes ranged from 25 to 200 \((n = 25, 50, 100, 150, 200, n < N)\), corresponding to a sampling fraction from 1.5% - 75%. In total, 15,120 scenarios were evaluated.

To compare the variance of the rate estimators, we recorded whether the variance of gender-adjusted rates was smaller than the variance of the crude rates. Ratios of variances of the adjusted rates (AR) to variances of the crude rates (CR) were calculated. The variance ratios were then plotted against percentage of males in population, by population smoking prevalence rate, population size and sample size. Thresholds of RRs, above which adjusted rates have smaller expected variances, were estimated graphically.

Results

We present results graphically in terms of a set of threshold relative risk of smoking by gender (Male/Female). The estimated thresholds for prevalence rate ratios are presented in Figure 1. Results for scenarios with population size of 3200 are not included due to their similarities to those with population size of 3200.
As shown in Figure 1, the variance reductions due to gender-adjustment depend on sample sizes, gender ratios, prevalence rate ratios, and overall prevalence rate of the population. Using the empirical estimators based on sample estimates of covariance between smoking status and male gender ($\hat{\sigma}_{xy}$), a reduction in the variance will occur only when the male-to-female prevalence rate ratios are sufficiently large, or the association between smoking status and the covariate are sufficiently strong.

In populations with balanced gender ratios and overall prevalence rates of 35%, the adjusted rates have smaller variances when RRs are above 1.6, 1.5, 1.4, 1.3 and 1.2 for sample sizes of 25, 50, 100, 150 and 200, respectively. In populations with unbalanced gender ratios, the RR thresholds are higher.

When the population prevalence rates are lower, for example, at 15% in gender balanced populations, the RR thresholds are much higher, i.e., 2.6, 2.1, 1.8, 1.6 and 1.5 for sample sizes of 25, 50, 100, 150 and 200, respectively.

The RR thresholds are much higher in populations with high male proportions, in particular, when sample sizes are under 150. For example, at 70% males in a population of size 1,600 with prevalence rate of 15%, the RR thresholds were above 4, 3, 2.3, 1.6, and 1.5 for sample sizes of 25, 50, 100, 150 and 200, respectively.

Sampling fractions influence the variance reduction as well, with higher sampling fraction related to slightly higher RR thresholds. For example, with a sample of size 150 from a population with balanced gender ratio and prevalence of 15%, the RR thresholds are 1.7, 1.55, 1.45 and 1.4 for populations of sizes 200, 400, 800 and 1,600 (corresponding to sampling fractions of 75%, 38%, 19%, and 9%). According to these data, in a study of a small town of 1600 adults, with a sample of 200, if the RR of
smoking with gender matches the Massachusetts population RR (1.13), gender-adjustment should not be used.
Discussion

In this analysis, we illustrate a danger in using adjustment for confounding. Although this study was limited to smoking and gender, the results have broad implications for other settings in epidemiological applications. The basic idea is that adjustments for confounding involve estimation of rates (or regression coefficients) from the sample. If the relationship between the confounder and the response variable is not strong, the added variability due to using estimated rates rather than true rates can overshadow the gain, usually viewed as resulting in a reduction in bias. The implication is that wider confidence intervals will result for the prevalence estimates if control for confounding is used. In fact, when the RR is lower than the thresholds indicated in Figure 1, gender-adjusted rates have precisions worse than the crude rates.

We simulated a large number of scenarios that are similar to common situations in epidemiologic sample surveys, where sample sizes are small and prevalence rates are relatively low. In certain situations, a much stronger relationship is needed for the adjustment for confounding to be warranted. For example, with a sample size of 100 and prevalence rate of 20%, the variance reduction will occur only if the prevalence rate ratio is greater than 1.7 in a population with balanced gender proportions, or over 2.5 in a population with 70% males and 30% females.

Covariate-adjusted rates, in theory, do lead to rate estimates that are more precise than crude rates when the regression coefficient $\beta$ or the relevant variance components ($\sigma_{xy}$ and $\sigma_x^2$) are known (Särndal et al. 1992). In practice, these parameters are rarely known, leading in some cases to more variable results. Small sample sizes ($n < 200$) are often encountered when analyzing small-scale sample surveys or in subgroup analysis of larger epidemiologic datasets. In such scenarios, the
simulation studies provide guidelines for whether or not adjustment is desirable. Based on analysis of a large number of scenarios simulated to mimic the practical situations, we have provided empirical guidelines with regard to situations where covariate-adjusted rates should not be used in place of crude rates. Such guidelines are presented in Figure 1 for many settings. A program (in Stata 9.0 \textit{(working in progress)}) is available on our web site that can be used to evaluate whether confounding should be controlled in other settings.

In this paper, we illustrated very simple scenarios where there is only one covariate (gender) with two classes (male versus female). More complicated yet common scenarios, such as those with multiple subgroups (e.g., race-ethnicity), or multiple categorical covariates (e.g., gender and race-ethnicity), or with mixture of categorical covariate (gender) and continuous covariate (age), should be further investigated. In particular, the impacts of the uncertainty of variance component estimation on the variances of covariate-adjusted rate estimators should be evaluated by Monte Carlo simulations. We are currently developing simulation results that address these settings.

The findings of our analysis may have important implications on the reporting methods of the public health sample survey data, such as BRFSS surveys, in particular, how the municipal (county, town or city)-level statistics should be reported. For example, among the 351 communities (towns or cities) of the Commonwealth of Massachusetts, 49 have a sample size $\geq 30$, and only 7 have a sample size $\geq 100$ in the 2003 BRFSS survey data. If the smoking prevalence rates of the 49 municipals with sample sizes $\geq 30$ are to be reported based on direct estimates from the sample, our empirical guidelines suggest that gender-adjusted rates should not be used in place of crude rates.
In summary, we caution about automatically including control for confounding without reviewing evidence of a relationship between the response variable and the confounder. Covariate-adjusted rates should not be used in all settings, and in particular, not when prevalence rates, prevalence rate ratios and sample sizes are small. In such settings, covariate-adjustment will lead to worse precision of the rate estimates and only an illusion of statistical control. We anticipate that this problem will be aggravated by inclusion of multiple covariates, but this problem warrants further investigation.
Reference


Legend

Figure 1 Relative risk threshold for variance reduction due to gender adjustment by prevalence rate, population and sample sizes