Using Logistic Regression to Estimate the Adjusted Attributable Risk of Low Birthweight in an Unmatched Case-Control Study

Charles Kooperberg[1] *and Diana B. Petitti*[2]

Other authors have shown how to estimate attributable risk based on stratification. In this paper, we show how to estimate adjusted attributable risks, standard errors, and confidence intervals from an unmatched case-control study that has population-based controls and uses the logistic regression model to estimate relative risk. We apply the method to data from a case-control study of low birthweight. The method is conceptually simple, has no assumptions beyond those of the logistic model, makes use of computer-intensive statistical techniques (the bootstrap), and extends to interactions. A Fortran computer program to carry out the computations is available from the authors upon request. (Epidemiology 1991;2;363–366)

Keywords: low birthweight, methods, attributable risk, bootstrap.

Several authors have discussed the problem of estimating the adjusted attributable risk based on data from case-control studies.[1-4] Whittemore[1,2] showed how to estimate attributable risk, adjusting for one dichotomous covariate. Bruzzi et al.[3] presented an approach adjusting for several factors simultaneously, based on estimates of relative risk derived from a logistic regression model, but they did not provide a way to estimate standard errors for the attributable risk estimates. Kuritz and Landis[4,5] presented a method for obtaining summary estimators, variances, and confidence intervals for attributable risk from both unmatched and matched case-control studies based on stratified analysis. They showed that, considering bias and coverage probability, their method for estimating variances and confidence intervals was superior to weighting the attributable risk estimates from each table by the inverse of its variance[6,7] and to weighting the attributable risk estimate from each table by the number of cases in the table.[1,2,8]

In this paper, we show how to estimate adjusted attributable risks, standard errors, and confidence intervals from an unmatched case-control study where controls are "population-based" (that is, sampled at random from the defined population from which cases also arose), using the logistic regression model to estimate relative risk. Estimating attributable risks, standard errors, and confidence intervals from the logistic model is of interest because this model is frequently used in epidemiologic studies, being particularly well-suited to analysis of data where covariates include measures that are both categorical and continuous.

We apply our method to analysis of data from a study examining the risk factors for low birthweight in an urban population.[9,10]

Methods

Estimation of the Adjusted Attributable Risk

All low-birthweight infants born during the study period were included in our sample, while the controls were chosen at random from noncases in the population, with sampling fraction p_1, so that the number of controls equaled the number of cases.

To obtain estimates for the attributable risk, a logistic model is fitted to the data. For each subject, there is a response variable y_j, which is 1 if subject j is a case and 0 if subject j is a control. There are data on a number of covariates x_{ij}. These variables are all dichotomous, that is, $x_{ij} = 1$ if subject j has risk factor i, and 0 otherwise.

The logistic model for the probability of low birthweight given the covariates is

$$P(y_j = 1 \mid x_j) = \frac{\exp\left(\alpha + \sum_{i=1}^{k} \beta_i x_{ij}\right)}{1 + \exp\left(\alpha + \sum_{i=1}^{k} \beta_i x_{ij}\right)}. \qquad (1)$$

The parameters α and β_i are estimated by maximum likelihood. To begin, consider the maximum likelihood estimates $\tilde{\alpha}$, $\tilde{\beta}_i$, $i = 1, \ldots k$, based on our data and the logistic model, but ignoring the fact that sampling took place. That is, we base $\tilde{\alpha}$ and $\tilde{\beta}_i$ on all cases and controls in the study and compute these estimates as if all low birthweight and all normal birthweight in the population

From [1]the Department of Statistics, University of California at Berkeley and [2]the Department of Family and Community Medicine, University of California at San Francisco. Address reprint requests to Diana B. Petitti, FCM, AC-9, Box 0900, San Francisco, CA 94143.

were included in the study. The maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}_i$, $i = 1, \ldots k$ of the parameters in Equation 1, taking the sampling into account, are then obtained as follows:

$$\hat{\alpha} = \tilde{\alpha} + \log(p_1)$$
$$\hat{\beta}_i = \tilde{\beta}_i \quad \text{for all } i,$$

following Breslow and Day.[11]

The expected number of cases in the population is obtained by adding up the individual probabilities of low birthweight, given the covariates:

$$C = \sum_j P(y_j = 1 \mid x_j).$$

The sum is not just over all subjects in the study, but over the entire population. For the controls, only a fraction p_1 of the eligible population was sampled, so the expected number of cases, taking sampling into account, can be estimated from the logistic model as

$$C = \sum_{y_j=1} \frac{\exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)} + \frac{1}{p_1} \sum_{y_j=0} \frac{\exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)}.$$

Here the controls are weighted by $1/p_1$.

To estimate the attributable risk for say, variable 1, we compute what would happen if nobody in the population has risk factor 1. Thus, the number of cases C expected if nobody in the population has risk factor 1, that is, $x_{1j} = 0$ for all j:

$$C_1 = \sum_{y_j=1} \frac{\exp\left(\hat{\alpha} + \sum_{i=2}^{k} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=2}^{k} \hat{\beta}_i x_{ij}\right)} + \frac{1}{p_1} \sum_{y_j=0} \frac{\exp\left(\hat{\alpha} + \sum_{i=2}^{k} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=2}^{k} \hat{\beta}_i x_{ij}\right)}.$$

The estimated attributable risk, AR, for the main effect of variable 1 is

$$\widehat{AR}_1 = \frac{C - C_1}{C} \times 100\%. \qquad (2)$$

Likewise, we can estimate the attributable risk for the remaining variables or for interactions between variables. To estimate, for example, $AR_{1,2}$, compute the number of cases $C_{1,2}$ expected if nobody had either risk factor 1 or risk factor 2.

Standard Errors and Confidence Intervals

We turn now to standard errors. There are no exact formulas for the standard error of the estimate (Equation 2) because of the complicated nonlinearities. Therefore, we use a method based upon the bootstrap (see Efron,[12,13] Freedman and Peters,[14] and Peters and Freedman[15]).

The basic idea is to see how good the estimates are in a situation where the true value of the parameter, say AR_1, is known. Since we do not know what AR_1 is, we construct data sets in a simulated world where we do know the parameter. The newly constructed data sets are identical to the original one, except for the response variable Y. The new Y*s are generated according to the probabilities specified by equation 1. The α and β in this equation are the estimates $\hat{\alpha}$ and $\hat{\beta}$. For each artificial data set, we estimate AR_1. Call these estimates $\widehat{AR}_1^*$. In our simulation, the truth, AR_1^*, is known: it is $\widehat{AR}_1^*$. Thus, from the difference between the $\widehat{AR}_1^*$ based upon the simulated data sets (the bootstrap estimates) and since $AR_1^* = \widehat{AR}_1^*$, we estimate the accuracy of $\widehat{AR}_1^*$. These estimates of accuracy of the bootstrap estimates are now used as an estimate of the accuracy of $\widehat{AR}_1^*$.

To generate artificial data sets exactly requires information on the entire population. This was not available because only a sample of mothers of normal infants was interviewed. Therefore, to get bootstrap estimates, we first reconstruct the population from which the cases arose. That is, all cases are in the population; all controls are replicated $1/p_1$ times (rounded to an integer) in the reconstructed population. Use of this replication leads to a negligible underestimation of the standard errors.

Next, to get one artificial data set Y*, we generate a pseudorandom number for every subject j in the reconstructed population. If this number is smaller than

$$\frac{\exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right)},$$

the jth subject will be labeled low birthweight, that is, $y^* = 1$; otherwise, it will be labeled normal birthweight, that is, $y^* = 0$. We use the same sampling scheme as was used in the study: All the low-birthweight infants are included in the data set, and the controls are sampled from the normal birthweights, so that the number of

TABLE 1. Attributable Risk (95% CI) of Low Birthweight for Six Variables in Blacks and Whites, Adjusting for All Other Variables in the Table

Variable	Whites			Blacks		
	AR*	SE*	95% CL*	AR	SE	95% CL
Age†						
≤17	0.8	0.8	−0.8, 2.4	2.0	1.8	−1.2, 5.5
35+	10.5	4.0	2.8, 17.9	1.8	1.9	−2.0, 5.3
Parity‡						
0	23.9	8.1	6.8, 38.1	12.6	4.9	2.9, 21.8
3+	−0.5	1.7	−4.0, 2.5	3.7	2.3	−1.0, 8.3
Low prepregnancy weight§	9.2	2.7	4.2, 14.6	19.3	2.5	14.4, 24.3
Drug use	−0.3	2.8	−6.5, 4.6			
Cocaine				8.2	1.9	4.2, 11.8
Other illegal‖				1.2	1.5	−2.2, 3.7
Heavy alcohol use¶	1.9	1.9	−2.2, 5.5	2.1	1.6	−1.7, 4.8
Smoking#	19.0	4.5	9.7, 26.9	31.3	4.2	23.1, 39.6

NOTE: Attributable risk is negative when the elimination of a factor would increase the number of low-birthweight infants, for example, when the factor "protects" from low birthweight.
*AR = attributable risk; SE = standard error; CL = confidence limits.
†Reference group is age 18–34.
‡Reference group is parity 1–2.
§Body mass index < 2.7 as measured in pounds and inches.
‖Except marijuana; includes heroin, amphetamines, PCP, angel dust, LSD.
¶3 or more drinks per day.
#Regular smoking.

controls equals the number of cases in this data set. We proceed by fitting the logistic model, computing $\hat{\alpha}^*$ and $\hat{\beta}^*$, and by estimating the attributable risk for this artificial sample, computing $\widehat{AR}_1^*$.

Now repeat this procedure many times (we used 1,000 repetitions), obtaining a large number of bootstrap estimates for AR_1^*. The standard deviation of the bootstrap estimates, $\widehat{AR}_1^*$, is the estimate for the standard error of the main effect of the attributable risk for variable 1; the sample 2.5th and 97.5th quantile of the bootstrap estimates form an approximate 95% confidence interval.

Essentially, the same method can be used for the attributable risk of interactions of variables, and it can easily be extended to include the situation where only the controls are not a probability sample from all eligible controls, but also if the cases are sampled from some larger set of eligible cases. It would also work for other models than the logistic model.

Application and Results

The details of procedures for recruiting and interviewing cases and controls have been described in detail elsewhere.[9,10] Briefly, we used information from birth certificates filed in Alameda County, California, to identify singleton infants without congenital anomalies born between January 1, 1987 and December 31, 1987 to white, non-Hispanic, and black residents of Alameda County. Cases weighed 500–2,499 grams, and all were targeted for interview. Controls were chosen at random from among infants with weights of 3,000 or more grams, the number of controls in each ethnic group being equal to the number of cases. The sampling fraction for white controls was 1/26.32; for black controls, it was 1/7.75. Two hundred and twenty-three white cases, 239 white controls, 377 black cases, and 389 black controls were successfully interviewed.

For blacks and whites separately, Table 1 shows estimates of the attributable risk of low birthweight for the main effect of six variables (age, parity, low prepregnancy weight, heavy use of alcohol, cigarette smoking, and use of cocaine and other illegal drugs). Standard errors and 95% confidence intervals are also shown. In both whites and blacks, after adjustment for all of the other variables, smoking accounted for the highest percentage of cases of low birthweight (AR% = 19, 95% CI: 10–27 for whites; AR% = 31, 95% CI: 23–40 for blacks). In blacks, after adjustment for all of the other variables including smoking and alcohol, use of cocaine accounted for 8% of cases of low birthweight (95% CI: 4–12).

Other models that we tried gave essentially the same fit and estimates for attributable risk for these six variables. The logistic model that included interactions also gave the same estimates.

These results suggest that elimination of substance abuse during pregnancy would prevent a substantial percentage of low-birthweight births in both whites and

blacks in this population, with smoking being numerically the most important contributor to the problem of low birthweight. The findings reinforce conclusions we made in two prior analyses based on data from this study that did not present adjusted attributable risk estimates.[9,10]

Discussion

As a tool for the program planner, attributable risk has important advantages over other measures of association derived from epidemiologic studies because it takes into account not only the strength of the factor's association with the disease or condition being studied, but also the prevalence of exposure. If the association of the factor with disease or condition is causal, attributable risk measures the percentage of all cases that would be prevented by eliminating the risk factor from the population. From the policy perspective, factors with high attributable risk should receive priority for preventive intervention, irrespective of the size of the relative risk of the disease or condition in the exposed.

A practical limitation of use of attributable risk for program planning has been the difficulty of estimating attributable risk after taking into account other factors known to be related to the disease. Kuritz and Landis[2-4] recently showed how to do this in a stratified analysis. The disadvantage of stratification is the need for a prohibitively large amount of data if one wants to take a number of variables into account at the same time. Using a model makes it possible to do computations with a considerably smaller amount of data. Modeling, however, has its risks. If the data do not fit the model, estimates for the attributable risk will be incorrect; so, of course, will be the standard errors and confidence intervals.

A computer program to carry out the computations, written in Fortran, is available from the authors upon request. Using the bootstrap to compute standard errors and confidence interval is highly computer intensive; it took 40 minutes CPU time on a Sparc-station to compute confidence intervals for the main effects of the attributable risk of low birthweight for blacks, using 1,000 bootstrap estimates. The computation of the estimates, which does not involve the bootstrap, was much faster. If only a PC is available, one could do a number of bootstrap estimates (say, 25) and compute standard errors using these estimates. Quantiles of the normal distribution are then used to construct approximate confidence bounds from the bootstrap estimates and standard errors.

The approach we present has some other important features. First, the only assumptions made are the standard assumptions behind a logistic model. Second, the method of obtaining estimates is conceptually simple. Third, the computation of confidence intervals and standard errors makes use of computer-intensive statistical techniques. Fourth, the method works not only for main effects, but it can easily be extended to include interactions of variables.

Acknowledgment

This research was supported by a grant (HD-19830) from the National Institute of Child Health and Human Development. David Freedman provided much helpful advice.

References

1. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. Stat Med 1982;1:229–243.
2. Whittemore AS. Estimating attributable risk from case-control studies. Am J Epidemiol 1983;117:76–85.
3. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 1985;122:904–914.
4. Kuritz SJ, Landis JR. Summary attributable risk estimation from unmatched case-control data. Stat Med 1988;7:507–517.
5. Kuritz SJ, Landis JR. Attributable risk estimation from matched case-control data. Biometrics 1988;44:355–367.
6. Walter SD. The estimation and interpretation of attributable risk in health research. Biometrics 1976;32:829–849.
7. Ejigou A. Estimation of attributable risk in the presence of confounding. Biometrical J 1979;21:155–165.
8. Walter SD. Calculation of attributable risks from epidemiological data. Int J Epidemiol 1978;7:175–182.
9. Petitti DB, Coleman D. Cocaine and the risk of low birth weight. Am J Public Health 1990;80:25–28.
10. Alameda County Low Birth Weight Study Group. Cigarette smoking and the risk of low birth weight: a comparison in black and white women. Epidemiology 1990;1:201–205.
11. Breslow NE, Day NE. Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies. Lyon: International Agency on Research on Cancer, 1980.
12. Efron B. Bootstrap methods: Another look at the jackknife. Ann Stat 1979;7:1–26.
13. Efron B. The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. Philadelphia: Society for Industrial and Applied Mathematics, 1982.
14. Freedman DA, Peters C. Bootstrapping a regression equation: Some empirical results. JASA 1984;79:97–106.
15. Peters SC, Freedman DA. Some notes on the bootstrap in regression problems. J Business Economic Stat 1984;2:406–410.