On a Closed-form Doubly Robust Estimator of the Adjusted Odds Ratio for a Binary Exposure

Eric J. Tchetgen Tchetgen

Departments of Epidemiology and Biostatistics,

Harvard University

Correspondence: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

This note concerns a goal common to many epidemiologic studies; mainly the estimation of the adjusted odds ratio of a binary exposure as it relates to the risk of a binary disease outcome. Because confounding bias is of serious concern in observational studies, investigators typically estimate an adjusted odds ratio in a multivariate logistic regression which conditions on a large number of potential confounders. It is well known that modeling error in specification of the confounders can lead to substantial bias in the adjusted odds ratio for exposure. As a remedy, Tchetgen Tchetgen et al (Biometrika 2010 vol. 97(1), pages 171-180) recently developed so-called doubly robust estimators of an adjusted odds ratio by carefully combining the standard logistic regression with the reverse regression analysis in which exposure is the dependent variable and both the outcome and confounders are the independent variables. Double robustness implies that only one of the two modeling strategies needs to be correct for valid inferences about the odds ratio parameter. This note aims primarily to introduce this recent methodology to the epidemiologic literature by presenting a simple closed-form doubly-robust estimator of the adjusted odds ratio for binary exposure. A SAS macro is given in an appendix to facilitate the use of the approach in routine epidemiologic practice and a simulated data example is also provided for the purpose of illustration.

Many epidemiologic studies aim to estimate using observational data, an adjusted odds ratio for a binary exposure A as it relates to the risk of a binary disease outcome Y, conditional on a moderate to large number of potential confounders L.[1,2] Logistic regression is widely used as the standard analytic tool for estimating the exposure-disease adjusted odds ratio which we denote $\exp(\psi^*)$, by fitting using maximum likelihood estimation, a working model

$$\text{logit}\{\Pr(Y = 1|A, L; \psi, \eta)\} = \eta_1 + \eta_2^T L + \psi A \tag{1}$$

for the conditional probability of the occurrence of the outcome given exposure level A and confounders L, where $\eta = (\eta_1, \eta_2)$. Throughout, we assume that as encoded in model (1), the effect of exposure is homogeneous on the odds ratio scale; that is we assume that for any individual with covariates L, the log-odds ratio for the exposure is equal to the constant ψ^*, the same for all values of L, so that ψ^* is the true value of ψ in model (1). It is important to note that, in the event that unmeasured confounding is present, ψ^* will generally fail to have a causal interpretation. However, it remains a well-defined summary measure under model (1) of the partial association between exposure and outcome after adjustment for the measured covariates L and it often constitutes the primary target of inference in epidemiological analyses. Our discussion concerns inference about the parameter ψ^* regardless of whether or not it has a meaningful causal interpretation, i.e. whether or not L is sufficiently rich to fully adjust for confounding.

Reverse regression and double robustness

Instead of the standard logistic regression described above, an alternative approach to estimate ψ^* less commonly used in epidemiologic practice, entails fitting a working logistic model for the reverse regression of A on (Y, L), such as

$$\text{logit}\{\Pr(A=1|Y,L;\psi,\alpha)\} = \alpha_1 + \alpha_2' L + \psi Y \tag{2}$$

where $\alpha = (\alpha_1, \alpha_2')'$. Crucially, the parameter ψ is shared between models (1) and (2) reflecting the following key property of odds ratios:

$$\begin{aligned} \exp(\psi^*) &= \frac{\Pr(Y=1|A=1,L)\Pr(Y=0|A=0,L)}{\Pr(Y=1|A=0,L)\Pr(Y=0|A=1,L)} \\ &= \frac{\Pr(A=1|Y=1,L)\Pr(A=0|Y=0,L)}{\Pr(A=1|Y=0,L)\Pr(A=0|Y=1,L)} \end{aligned} \tag{3}$$

therefore, the models (1) and (2) represent two genuine (i.e. independent) opportunities to correctly estimate the exposure effect; in the sense that the two working models are perfectly compatible for any value of α and η. However, even if model (3) is known to hold, in practice, it is generally impossible to know with certainty which, if any, of regression models (1) and (2) is correctly specified. The best that one can hope for is that these two distinct estimation strategies can be combined into a single overall strategy which is guaranteed in large samples, to deliver a correct estimate of ψ^* provided that one of the two models (1) and (2) is correctly specified, without necessarily knowing which is correct. Tchetgen Tchetgen and colleagues have recently developed a large class of estimators with precisely this desirable property. Stated more precisely, the methods of Tchetgen Tchetgen et al[3] yield a specific class of estimators of ψ^* that are asymptotically unbiased if one, but not necessarily both of the following is true:

(i) the working model $\Pr(A=1|Y=0,L;\alpha)$ is correctly specified even if $\Pr(Y=1|A=0,L;\eta)$ is incorrectly specified,

(ii) the working model $\Pr(Y=1|A=0,L;\eta)$ is correctly specified even if the working model

$\Pr(A = 1|Y = 0, L; \alpha)$ is incorrectly specified.

Unfortunately, the estimator of Tchetgen Tchetgen et al[3] is not easily obtained without special software, thus seriously impeding its routine use. This computational challenge inspired the SAS macro of Tchetgen Tchetgen and Rotnitzky[4] which implements the optimal doubly-robust estimator via an iterative procedure which they called the *ProRetroSpec algorithm*. Here, the above computational challenge is further addressed in the simple case of a binary exposure, and a simple closed-form doubly robust estimator is provided which is easy to compute using standard software.

Suppose that independent and identically distributed data on (A, Y, L) is observed on n individuals in whom the homogeneous odds ratio model (3) is known to hold. Let $\widehat{P}_i = \Pr(A_i = 1|Y = 0, L_i; \widehat{\alpha})$ where $\widehat{\alpha}$ denotes the maximum likelihood estimator of α using data on non-cases only (i.e. with $Y = 0$) under model (2); similarly, let $\widehat{B}_i = \Pr(Y_i = 1|A_i = 0, L; \widehat{\eta})$ where $\widehat{\eta}$ denotes the maximum likelihood estimator of η using data on the unexposed subsample only (i.e. with $A = 0$) under model (1). Let $W_i = w(L_i)$ denote a user-specified function of L_i.A simple closed-form estimator of ψ^* is defined as $\widehat{\psi}(w) =$:

$$\log \frac{\sum_i W_i \left\{ A_i Y_i \left(1 - \widehat{P}_i\right) \left(1 - \widehat{B}_i\right) \right\}}{\sum_i W_i \left\{ A_i (1 - Y_i) \left(1 - \widehat{P}_i\right) \widehat{B}_i + (1 - A_i) Y_i \widehat{P}_i \left(1 - \widehat{B}_i\right) - (1 - A_i)(1 - Y_i) \left(1 - \widehat{P}_i\right) \left(1 - \widehat{B}_i\right) \right\}}$$

For a fixed choice (possibly estimated) of $w(\cdot)$, the resulting estimator $\widehat{\psi}(w)$ is doubly-robust and thus converges to ψ^* (in probability) with increasing sample size, provided that either condition (i) or (ii) holds, but not necessarily both. A formal argument establishing double robustness of $\widehat{\psi}(w)$ is given in an online appendix for completeness. The choice for the weight function $w(\cdot)$ only affects efficiency, and the optimal choice of the weight $w(\cdot)$ is easily inferred from a result in Tchetgen Tchetgen et al[3], and is beyond the scope of this note. Sample SAS code for the simple

estimator $\widehat{\psi} = \widehat{\psi}(1)$ is provided in the online appendix, along with code for computing standard errors that correctly account for the variation in $\widehat{P}_i$ and $\widehat{B}_i$.

A data illustration

We briefly illustrate the application of the doubly robust macro in a simulated data set. For this purpose, we generated independent and identically distributed data on 2000 individuals under models (1) and (2) with covariates $L = (L_1, L_2, L_3 = L_1 \times L_2)$, where (L_1, L_2) are independent standard normal; $\psi^* = 0.5$, and regression coefficients $\alpha = (0.5, -0.3, 0.5, -0.5)'$, $\eta = (0.5, 0.5, -0.5, -0.5)$. In order to illustrate the presence of bias under model mis-specification of (1), we obtained estimates of ψ^* using the standard (prospective) logistic regression (1) both with and without the interaction term L_3 which produced $\widehat{\psi}_p^1 = 0.56$ (s.e.=0.11) and $\widehat{\psi}_p^2 = 0.76$ (0.10) respectively. Similarly, we obtained estimates of ψ^* in the reverse regression (2) both with and without a term for the interaction L_3 which gave $\widehat{\psi}_r^1 = 0.53$ (s.e.=0.11) and $\widehat{\psi}_r^2 = 0.75$ (s.e.=0.10) respectively. Finally, recall that the doubly robust estimator uses both models (1) and (2) to estimate ψ^*, and we aim to show that only one model needs to be correct for valid inference. Thus, we obtained three doubly robust estimators. the first estimator $\widehat{\psi}_{dr}^1 = 0.54$ (s.e.=0.11) used correct models for (1) and (2). The second estimator $\widehat{\psi}_{dr}^2 = 0.55$ (s.e.=0.10) used the correct outcome regression model (1) but mis-specified the exposure model (2) by leaving out the interaction, while $\widehat{\psi}_{dr}^3 = 0.56$ (s.e.=0.10) was obtained under the reversed situation, that is by mis-specifying (1) but with a correct specification of the model (2).

As one would expect, both logistic models (1) and (2) are consistent under correct model specification and are severely biased when the interaction between confounders is omitted. This example nicely illustrates the doubly robust property, since the doubly robust estimator is consistent in the absence of modeling error, and it agrees with the outcome regression estimate $\widehat{\psi}_p^1$ when

the latter is consistent, and it similarly agrees with the reverse regression $\widehat{\psi}_r^1$ when the latter is consistent. Therefore, this example provides an empirical confirmation of the theoretical property that the doubly robust approach is consistent as long as at least one of models (1) and (2) is correct, without knowing which one holds. For completeness, we provide an example of the macro call used to estimate $\widehat{\psi}_{dr}^2$:

```
%dr_odds(data=_cero_,y=outcome, a=exposure, oddsY=l1 l2, oddsA=l1 l2 l3,sample=2000) ;
```

where the first argument to the macro gives the name of the data set which contained the variables (outcome,exposure, l1,l2,l3); the second argument specifies the outcome variable; the third argument specifies the exposure variable; the fourth argument lists the covariates to be included in the outcome regression model (1) ; the fifth argument lists the covariates to be included in the exposure model (2); and the final argument sets the sample size.

Final remarks

In closing, we note that the doubly robust methodology described herein generalizes to polytomous and continuous exposure, and can also incorporate effect modification of the exposure-outcome odds ratio by components of L; although closed-form estimators as obtained herein are generally not available in such more general settings.[3] Additionally, it is worth noting that because the odds ratio effect measure generally remains identified under a (matched) case-control design, the doubly robust methodology described herein equally applies under such outcome dependent sampling designs.[4]

References

[1] Rothman K and Greenland S. Modern Epidemiology. Third Edition. 2008. Lippincott Williams and Wilkins.Philadelphia

[2] Breslow NE, Day NE. Statistical Methods in cancer research.Vol I:The analysis of case-control data. Lyon: IARC.1980.

[3] Tchetgen Tchetgen E, Robins J, Rotnitzy A. On doubly robust estimation in a semi-parametric odds ratio model. Biometrika 2010 vol. 97(1), pages 171-180.

[4] Tchetgen Tchetgen EJ, Rotnitzky A: Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. Stat Med; 2011 Feb 20;30(4):335-47.

[5] SAS Institute, Inc. SAS/STAT User's Guide, Version 8 Cary, NC. 1999.

APPENDIX

Proof of double robustness:

The proof follows from noting that $\widehat{\psi}(w)$ solves the equation:

$$
\begin{aligned}
0 &= \sum_i U_i\left(\widehat{\psi}, \widehat{\alpha}, \widehat{\eta}\right) \\
&= \sum_i W_i \left\{\left(A_i - \widehat{P}_i\right)\left(Y_i - \widehat{B}_i\right)\exp\left(-\widehat{\psi}(w) A_i Y_i\right)\right\}
\end{aligned}
$$

therefore $\widehat{\psi}(w)$ converges to ψ^* only if ψ^* satisfies the population equation

$$
0 = E\left[W\left\{\left(A - P^{\#}\right)\left(Y - B^{\#}\right)\exp\left(-\psi^* AY\right)\right\}\right]
$$

where $\left(P^{\#}, B^{\#}\right)$ is the probability limit of $\left(\widehat{P}, \widehat{B}\right)$,and E indicates expectation. We show that the latter property holds provided that either $P^{\#} = \Pr(A|Y=0, L)$ or $B^{\#} = \Pr(Y|A=0, L)$ but both do not necessarily hold. consider the case where $P^{\#} = \Pr(A|Y=0, L)$, then note that

$$
f(A=a|Y, L) = \frac{f(A=a|Y=0, L)\exp\left(\psi^* aY\right)}{\sum_{a^*} f(A=a^*|Y=0, L)\exp\left(\psi^* a^* Y\right)}
$$

and

$$
\begin{aligned}
& E\left[W\left\{(A-P)\left(Y-B^{\#}\right)\exp\left(-\psi^{*}AY\right)\right\}|Y,L\right] \\
&= \sum_{a}\left[W\left\{(a-P)\left(Y-B^{\#}\right)f(A=a|Y,L)\exp\left(-\psi^{*}0Y\right)\right\}\right] \\
&= \sum_{a}\left[W\left\{(a-P)\left(Y-B^{\#}\right)\frac{f(A=a|Y=0,L)\exp\left(\psi^{*}aY\right)\exp\left(-\psi^{*}aY\right)}{\sum_{a^{*}}f(A=a^{*}|Y=0,L)\exp\left(\psi^{*}a^{*}Y\right)}\right\}\right] \\
&= \sum_{a}\left[W\left\{(a-P)\left(Y-B^{\#}\right)\frac{f(A=a|Y=0,L)}{\sum_{a^{*}}f(A=a^{*}|Y=0,L)\exp\left(\psi^{*}a^{*}Y\right)}\right\}\right] \\
&= 0
\end{aligned}
$$

since $\sum_a (a-P)\, f(A=a|Y=0,L) = \sum_a (a - f(A=1|Y=0,L))\, f(A=a|Y=0,L) = 0$. The proof is completed by a symmetric argument for the other case.

Asymptotic variance formula:

We first note that $\widehat{\alpha}$ solves the score equation

$$
\begin{aligned}
0 &= \sum_{i} S_{\alpha,i}\left(\alpha\right) \\
&= \sum_{i}(1-Y_i)(A_i - \Pr(A=1|Y=0,L;\alpha)[1,L]'
\end{aligned}
$$

and $\widehat{\eta}$ solves the score equation

$$
\begin{aligned}
0 &= \sum_{i} S_{\eta,i}\left(\eta\right) \\
&= \sum_{i}(1-A_i)(Y_i - \Pr(Y=1|A=0,L;\eta)[1,L]'
\end{aligned}
$$

The large sample variance of $\widehat{\psi}\left(w\right)$ can be derived by a standard Taylor expansion and is consis-

tently estimated by:

$$\frac{\sum_i \left\{U_i\left(\psi,\widehat{\alpha},\widehat{\eta}\right)+Q_i+R_i\right\}^2}{\left\{\sum_i \frac{\partial U_i(\psi,\widehat{\alpha},\widehat{\eta})}{\partial\psi}|_{\psi=\widehat{\psi}}\right\}^2}$$

where

$$Q_i=-\left\{\sum_i \frac{\partial U_i\left(\widehat{\psi},\alpha,\widehat{\eta}\right)}{\partial\alpha}|_{\alpha=\widehat{\alpha}}\right\}\left\{\sum_i \frac{\partial S_{\alpha,i}\left(\alpha\right)}{\partial\alpha}|_{\alpha=\widehat{\alpha}}\right\}^{-1} S_{\alpha,i}\left(\widehat{\alpha}\right)$$

$$R_i=-\left\{\sum_i \frac{\partial U_i\left(\widehat{\psi},\widehat{\alpha},\eta\right)}{\partial\eta}|_{\eta=\widehat{\eta}}\right\}\left\{\sum_i \frac{\partial S_{\eta,i}\left(\eta\right)}{\partial\eta}|_{\eta=\widehat{\eta}}\right\}^{-1} S_{\eta,i}\left(\widehat{\eta}\right)$$

We emphasize that this variance estimator is nonparametric in the sense that it converges to the asymptotic variance of $\widehat{\psi}$ irrespective of whether any of $\left(\widehat{P},\widehat{B}\right)$ is consistent.

Sample SAS Code

```
/*
Input to the macro:
data=Name of data set;
y=name of binary outcome variable;
a=name of binary exposure variable;
oddsY = list of variables to include in the outcome regression model,
            e.g."oddsY=gender age education" ;
oddsA=list of variables to include in the exposure regression model,
            e.g "oddsA=gender age education age*gender" ;
sample= sample size;
 Output of the macro:
psi=doubly robust estimator of the log-odds ratio
se_dr =standard error of doubly robust estimator
*/
%macro dr_odds(data= ,y=, a=, oddsY=, oddsA=,sample=) ;
data cero_bis;set &data; retain id 0;
id=id+1;
y_bis=&y;
a_bis=&a;
if &a=1 then y_bis=.;
if &y=1 then a_bis=.;
int=1;
```

```
run;

proc logistic data=cero_bis descending;

model y_bis=&oddsY;

output out=predy_bis p=py;

run;

proc logistic data =cero_bis descending;

model a_bis=&oddsA;

output out =preda_bis p=pa;

run;

data dr_estimate; merge predy_bis preda_bis; by id; retain num denum 0;

num=num+&a*&y*(1-py)*(1-pa);

denum=denum+(1-&a)*&y*(1-py)*pa+&a*(1-&y)*py*(1-pa)-(1-&a)*(1-&y)*pa*py;

if id=&sample then do; psi_dr=log(num/denum);output; end;

run;

proc iml;

use cero_bis ;

read all var {id} into id;

read all var {int} into int;

read all var {int &oddsY} into l_y ;

read all var {int &oddsA} into l_a;

read all var {&y} into y;

read all var {&a} into a;

use predy_bis;

read all var {py} into py;
```

```
use preda_bis;

read all var {pa} into pa;

use dr_estimate;

read all var {psi_dr} into psi;

U= (y-py)#exp(-psi*a#y)#(a-pa);

bread= 1/(((y-py)#exp(-psi*a#y)#(a-pa)#a#y)[+]);

cor_y=-t(py#(1-py) #exp(-psi*a#y)#(a-pa))*L_y;

score_y=(t((1-A)#L_y#py#(1-py))*(L_y#(1-A)))*t((y-py)#(1-A)#L_y);

cor_a=-t(pa#(1-pa) #exp(-psi*a#y)#(y-py))*L_a;

score_a=(t((1-y)#L_a#pa#(1-pa))*(L_a#(1-y)))*t((a-pa)#(1-y)#L_a);

meat_U=U+t(cor_y*score_y)+t(cor_a*score_a);

meat =t(u)*u;

se_dr= (bread*meat*bread)**.5;

create var from se_dr;

append from se_dr ;

print psi se_dr;

quit;

%mend;
```