Doubly robust estimation of causal effects

Web Appendices

Web Appendix 1: Demonstration of the Doubly Robust Property

A close examination of the statistical expression for the doubly robust estimator provides an intuitive illustration of the doubly robust property. We have adapted and expanded the proof given by Tsiatis (p148-149, (1)) to make it more accessible to non-statisticians. Equations have been included, but the text that accompanies them is non-technical. We recommend Bang & Robins (2) as an excellent intermediate reference and Tsiatis (1) or van der Laan and Robins (3) for an in-depth theoretical treatment of doubly robust methods.

Suppose we are interested in the causal effect of an exposure X (taking values 1 or 0 indicating presence or absence) on an outcome Y. Using a counterfactual framework, we say that $Y_{X=1}$ and $Y_{X=0}$ are the potential outcomes that would be observed in the presence and absence of the exposure, respectively (4). In addition, we have measured various baseline covariates (**Z**) that may be causally related to exposure and/or the outcome. All of these variables are further subscripted by *i* for individuals *i*=1, . . ., *n*. For illustration, we consider estimation of the difference in means due to exposure or, in other words, the mean response if everyone in the population were to be exposed ($E(Y_{X=1})$) minus the mean response if everyone were to remain unexposed ($E(Y_{X=0})$). One could similarly construct a relative effect measure using $E(Y_{X=1})$ / $E(Y_{X=0})$.

$$\begin{aligned}\hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}&\left[\frac{X_iY_i}{e(Z_i,\hat{\beta})}-\frac{\left\{X_i-e(Z_i,\hat{\beta})\right\}}{e(Z_i,\hat{\beta})}m_1(Z_i,\hat{\alpha}_1)\right] \\ -n^{-1}\sum_{i=1}^{n}&\left[\frac{(1-X_i)Y_i}{1-e(Z_i,\hat{\beta})}+\frac{\left\{X_i-e(Z_i,\hat{\beta})\right\}}{1-e(Z_i,\hat{\beta})}m_0(Z_i,\hat{\alpha}_0)\right]\end{aligned} \qquad \text{(eq 1)}$$

$$= \hat{\mu}_{1,DR} - \hat{\mu}_{0,DR} \qquad \text{(eq 2)}$$

In (eq 1) for the estimated effect of exposure $(\hat{\Delta}_{DR})$, the first term in each average is an inverse probability weighted estimator for $E(Y_{X=1})$ or $E(Y_{X=0})$, respectively. The second term is the "augmentation" that serves both to increase efficiency and support the doubly robust property. In (eq 2) for the mean difference due to exposure, $\hat{\mu}_{1,DR}$ estimates $E(Y_{X=1})$ and $\hat{\mu}_{0,DR}$ estimates $E(Y_{X=0})$.

The postulated model for the true PS is represented as *e*($\mathbf{Z}_i$,*β*). The expressions *m₀*($\mathbf{Z}_i$,*α₀*) and *m₁*($\mathbf{Z}_i$,*α₁*) are postulated outcome regression models for the true relations between the vector of covariates and the outcome within the unexposed and exposed, respectively. Here, $\hat{\beta}$, $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are estimates for the parameters β, α_0 and α_1 in the postulated models. The PS is estimated by substituting the estimate for $\hat{\beta}$ obtained by logistic regression. Similarly, m_0 and m_1 are estimated by substituting the estimates for $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from the outcome regression models.

$$n^{-1}\sum_{i=1}^{n}\left[\frac{X_iY_i}{e(Z_i,\hat{\beta})}-\frac{\left\{X_i-e(Z_i,\hat{\beta})\right\}}{e(Z_i,\hat{\beta})}m_1(Z_i,\hat{\alpha}_1)\right] \qquad \text{(eq 3)}$$

$$\approx E(Y_{X=1})+E\left[\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right] \qquad \text{(eq 4)}$$

To demonstrate the doubly robust property, we focus on the estimator for the average response in the presence of exposure, E($Y_{X=1}$), given by $\hat{\mu}_{1,DR}$, (eq 3) [first line of (eq 1)]. When *n* is large, the sample average (eq 3) estimates the population average (eq 4). The first term, E($Y_{X=1}$), is the average response with exposure. If the second term in (eq 4) reduces to zero, the entire quantity (eq 4) will estimate the average outcome with exposure. We present two scenarios: 1) a correct PS model but incorrect outcome regression model and 2) a correct outcome regression model but incorrect PS model. In each, we describe how the second term in (eq 4) reduces to zero.

First, consider the situation where the postulated PS model e(Z,β) is correct but the postulated outcome regression model m_1(Z,α_1) is not. That is e(Z,β)=e(Z)=E(X|**Z**) but m_1(Z,α_1)≠E(Y|X=1,**Z**). In the event that we specify the correct model for the PS, we can substitute e(**Z**) for e(Z,β) but the outcome regression model, having been misspecified, does not estimate E(Y|X=1,**Z**) and so we cannot make this substitution between equations (eq 5) and (eq 6).

$$E\left[\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right] \qquad \text{(eq 5)}$$

$$E\left[\frac{\{X-e(Z)\}}{e(Z)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right] \qquad \text{(eq 6)}$$

Nonetheless, when we manipulate (eq 6) algebraically (eq 7- eq 10) and invoke the exchangeability assumption (eq 10- eq 11), it reduces to zero.

$$E\left(E\left[\frac{\{X-e(Z)\}}{e(Z)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\middle|Y_{X=1},Z\right]\right) \qquad \text{(eq 7)}$$

$$=E\left(E\left[\frac{\{X-e(Z)\}}{e(Z)}\middle|Y_{X=1},Z\right]\{Y_{X=1}-m_1(Z,\alpha_1)\}\right) \qquad \text{(eq 8)}$$

$$=E\left(\frac{\{E(X\mid Y_{X=1},Z)-e(Z)\}}{e(Z)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right) \qquad \text{(eq 9)}$$

$$=E\left(\frac{\{E(X\mid Z)-e(Z)\}}{e(Z)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right) \qquad \text{(eq 10)}$$

$$=E\left(\frac{\{e(Z)-e(Z)\}}{e(Z)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right) \qquad \text{(eq 11)}$$

$$\begin{aligned}&=E\left(\{0\}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right)\\&=E(0)\end{aligned} \qquad \text{(eq 12)}$$

Therefore, even if the postulated outcome regression model is incorrect, $\hat{\mu}_{1,DR}$ estimates E($Y_{X=1}$) and similarly $\hat{\mu}_{0,DR}$ estimates E($Y_{X=0}$) such that the difference or ratio estimates the average causal effect of exposure.

Next, we consider the situation in which the outcome regression model is correct but the PS model is not. That is $m_1(Z,\alpha_1)$=E(Y|X=1,**Z**) but e(Z,β)≠e(**Z**)≠E(X|**Z**). In this instance, the second term in equation (eq 4) for $\hat{\mu}_{1,DR}$ (eq 13) can be rewritten as shown in (eq 14).

$$E\left[\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{Y_{X=1}-m_1(Z,\alpha_1)\}\right] \qquad \text{(eq 13)}$$

$$E\left[\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{Y_{X=1}-E(Y\mid X=1,Z)\}\right] \qquad \text{(eq 14)}$$

This term reduces to zero by manipulating the equation algebraically (eq 15- eq 17) and invoking the exchangeability assumption (eq 17- eq 18).

$$E\left(\left[\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{Y_{X=1}-E(Y\mid X=1,Z)\}\middle| X,Z\right]\right) \qquad \text{(eq 15)}$$

$$=E\left(\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}E\left[\{Y_{X=1}-E(Y\mid X=1,Z)\}\middle| X,Z\right]\right) \qquad \text{(eq 16)}$$

$$=E\left(\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{E(Y_{X=1}\mid X,Z)-E(Y\mid X=1,Z)\}\right) \qquad \text{(eq 17)}$$

$$=E\left(\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{E(Y_{X=1}\mid Z)-E(Y_{X=1}\mid Z)\}\right) \qquad \text{(eq 18)}$$

$$\begin{aligned}&=E\left(\frac{\{X-e(Z,\beta)\}}{e(Z,\beta)}\{0\}\right)\\&=E(0)\end{aligned} \qquad \text{(eq 19)}$$

Thus, equation (4) estimates the quantity of interest, E($Y_{X=1}$), even though the PS model was misspecified. As before, $\hat{\mu}_{1,DR}$ estimates E($Y_{X=1}$) and similarly $\hat{\mu}_{0,DR}$ estimates E($Y_{X=0}$) such that the difference or ratio estimates the average causal effect of exposure.

$$\hat{\Delta}_{DR}=\left[E(Y_{X=1})+augmentation\right]-\left[E(Y_{X=0})+augmentation\right] \qquad \text{(eq 20)}$$

$$\hat{\Delta}_{DR}=E(Y_{X=1})-E(Y_{X=0}) \qquad \text{(eq 21)}$$

Web Appendix 2: SAS Macro for Doubly Robust Estimation

The SAS macro described here uses the semi-parametric locally efficient augmented IPW estimator (5) that Scharfstein et al (6) noted is doubly robust. The macro can be downloaded along with documentation and the sample data from http://www.unc.edu/~mfunk/dr/.

Specifying the models

The macro requires the analyst to specify a model for the relationship between exposure and confounders (the weight model) as well as the models for the relationship between the outcome and confounders within strata of exposure. The macro uses the estimated parameters from all three models to calculate the estimated effect of exposure.

After loading the macro itself in SAS, the following three lines of code would provide the necessary parameters to the macro for estimating the effect of the exposure on the outcome:

```
%dr(%str(options data=sample desc;
      wtmodel e=x1 x2 x3   / method=dr dist=bin;
      model o=x1 x2 x3 x12 / dist=bin;));
```

The first line includes the necessary code for invoking the macro followed by the usual *data=* option for identifying the SAS dataset and the *descending* option to model the probability that the exposure=1 in the weight model and that the outcome=1 in the outcome regression model. The second line specifies the weight model (or propensity score model) with options indicating that the DR method is to be used (*method=dr*), that the distribution of the exposure is binary (*dist=bin*) and requesting that the propensity score curves stratified by exposure group be displayed (*showcurves*). The third line specifies the covariates to be used in the outcome regression models within each exposure group. In addition to the three covariates (Z_1, Z_2, and Z_3), we have include an interaction term between Z_1 and Z_2 (Z_{12}) that was created in a previous data step. Because the outcome regression models are conducted within exposure groups, the exposure variable itself is not listed as an independent variable here. The model statement also includes an option indicating that the outcome has a binomial distribution (*dist=bin*) which results in the use of a logistic model for estimating the effects of the covariates on the outcome.

Alternatively, the outcome could be a continuous variable that is normally distributed (*dist=n*) for which we use a set of linear regression models within each exposure group.

Output

The output of this macro includes the usual information on the three component regression models – the weight model as well as the two outcome regression models. The outcome regression models are inherently somewhat flexible in that they allow the parameter estimates for the effect of each confounder to vary by exposure group (equivalent to placing interaction terms into a model between the exposure and every covariate). For instance, if BMI was strongly associated with the outcome among the unexposed but not among the exposed, the outcome regression models within exposure group would estimate different values for those two parameters. The usual SAS output for each outcome regression model is provided so that the analyst can identify circumstances where this may be the case.

The macro incorporates diagnostic information about the weight model to assist the analyst in evaluating the appropriateness of the model and the resulting weights. The descriptive information (n, mean, standard deviation, min, max) for the propensity scores (*ps*) stratified by exposure group (*X*) is important to review. The degree of overlap of the propensity score values can be inspected visually in the histograms stratified by exposure group. When all of the confounders are dichotomous (as in this example), this leads to relatively disjoint propensity score 'curves', but nonetheless there is reasonable overlap with some unexposed individuals available to represent the outcomes of all exposed individuals and vice versa. There are no combinations of covariates that lead to a p(X=1) or p(X=0) equal to zero. This is known as the positivity assumption and is required for propensity score and IPW methods to be valid (7).

Effect estimates

The estimates of the average causal effect of the exposure appear as the final component of the output. In this example, the proportion of the population that is estimated to experience the outcome under no exposure (DR_0) is the same as the proportion that is estimated to experience the outcome with exposure (DR_1), or 0.22, indicating no effect of exposure on the outcome (deltaDR=0). The relative effect estimates are also null. (Running this code against the sample dataset provided online should produce the same numeric results.)

```
                              Primary Results

                                      Lower      Upper                   Prob
     Statistic    Estimate      SE*    95% CL     95% CL      Chi-sq      Chi-sq

     DR 0          0.22000     .          .          .              .     .
     DR 1          0.22000     .          .          .              .     .
     Delta DR      0.00000   0.009537   -0.01869   0.01869   1.5709E-16   1.00000
     Log RR        0.00000   0.043352   -0.08497   0.08497    1.572E-16   1.00000
     Log OR        0.00000   0.055579   -0.10893   0.10893    1.572E-16   1.00000
     RR            1.00000   NA          0.89679   1.11509    1.572E-16   1.00000
     OR            1.00000   NA          2.45172   3.04984    1.572E-16   1.00000

*NOTE: Estimated standard errors assume all models are correctly specified
```

These estimates can be interpreted as the difference or relative effect due to exposure were the entire population to have been exposed versus unexposed. In the case of a continuous outcome, the effect of interest is generally the absolute difference in the outcome (mean difference or *deltadr*). In the case of a dichotomous outcome, the effect of interest might be on the absolute scale (risk difference or *deltadr*) or on the relative scale (relative risk or odds ratio). All three are provided for a dichotomous outcome.

Bootstrapped standard errors and confidence intervals

As noted above, the estimated standard errors assume that both models have been correctly specified and are, therefore, not doubly robust. Thus, we strongly recommend the use of bootstrapping to obtain appropriate standard errors and confidence intervals for these estimates. The following sample code runs the same analysis as above with the addition of bootstrapping.

```
%dr(%str(options data=sample desc bootstrap=1000 alpha=0.05 bootout=bs_results;
      wtmodel e=x1 x2 x3   / method=dr dist=bin;
      model o=x1 x2 x3 x12 / dist=bin;));
```

Bootstrap=n requests that bootstrapped standard errors and confidence intervals be estimated based on *n* complete resamples of the data. Alpha=0.05 indicates that the confidence limits and p values should be based on a two-sided alpha of 0.05. The results from the analysis of the bootstrapped resamples can be saved for further examination by specifying the bootout=<*dataset_name*> option.

The output from this analysis appears below. Note that the numerical results obtained by running this code against the sample dataset online may differ slightly from those shown due to variability in bootstrapped resamples.

```
                      Means and Empirical Confidence Limits
                          Bootstrapped Iterations=1000

                           Empirical      Lower      Upper                   Prob
  Statistic    Estimate       SE         95% CL     95% CL      Chi-sq      Chi-sq

  DR 0          0.21973    0.006746        .          .               .      .
  DR 1          0.21982    0.006478        .          .               .      .
  Delta DR      0.00009    0.009185    -0.01800    0.01800    1.6937E-16    1.00000
  Log RR        0.00047    0.041792    -0.08191    0.08191    1.6916E-16    1.00000
  Log OR        0.00058    0.053566    -0.10499    0.10499    1.6924E-16    1.00000
  RR            1.00047    NA           0.90034    1.11070    1.6924E-16    1.00000
  OR            1.00058    NA           2.46043    3.03647    1.6924E-16    1.00000


                  Medians and Percentile-Based Confidence Limits
                          Bootstrapped Iterations=1000

                                         Lower      Upper
                  Statistic    Estimate     95% CL     95% CL

                  DR 0          0.21958    0.20763    0.23378
                  DR 1          0.21987    0.20726    0.23259
                  Delta DR     -0.00002   -0.01824    0.01804
                  Log RR       -0.00008   -0.08225    0.08095
                  Log OR       -0.00010   -0.10581    0.10436
                  RR            0.99992    0.92104    1.08432
                  OR            0.99990    0.89959    1.11000
```

Web Appendix 3: Data Generation for Monte Carlo Simulations

We generated three independent variables (Z_1, Z_2, Z_3). Z_1 and Z_2 were both normally distributed with mean=0 and standard deviation=1. Z_3 was a dichotomous variable generated by drawing a random number from a uniform distribution between 0-1 and assigning the value of Z_3 to 1 if the random number was <=0.3 and otherwise setting it to 0. With these independent variables defined, we then generated the exposure (X) as a function of these three variables. The true propensity score model was $P\{X=1|\mathbf{Z}]=(1+\exp\{-(\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 Z_3)\})^{-1}$ where $\beta_0=1.5$, $\beta_1=1$, $\beta_2=-2$, and $\beta_3=1$. In order to dichotomize the exposure variable, SAS drew a random number R from the uniform distribution between 0-1. If $[p(X=1|\mathbf{Z}_i)+R]$ was <0.91, X was set to 1. Otherwise, it was set to 0. The resulting exposure variable had a $p(X=1) = 0.2$. The outcome was a continuous variable (Y), generated as a function of Z_1 and Z_3, but not X or Z_2. Specifically, $E(Y)=\beta_1 Z_1+\beta_3 Z_3+\beta_4 Z_4$ where Z_4 was a randomly drawn number from a normal distribution with mean 0 and standard deviation of 1, $\beta_1=1$, $\beta_3=1$, *and* $\beta_4=2$. The resulting outcome variable had a mean of 0.3 and standard deviation of 2.3.

	Propensity score model	**Outcome regression models**
	$\text{Logit}P(X=1|\mathbf{Z})=$	$E(Y_{X=1})=$ & $E(Y_{X=0})=$
True models	$1.5+Z_1-2*Z_2+Z_3$	$Z_1+Z_3+2*Z_4$
Scenario 1	$\beta_0+\beta_1 Z_1+\beta_3 Z_3$	$\beta_0+\beta_1 Z_1+\beta_3 Z_3$
Scenario 2	$\beta_0+\beta_1 Z_1$	$\beta_0+\beta_1 Z_1+\beta_3 Z_3$
Scenario 3	$\beta_0+\beta_1 Z_1+\beta_3 Z_3$	$\beta_0+\beta_1 Z_1$

Web Appendix References

1. Tsiatis AA. Semiparametric Theory and Missing Data. New York: Springer, 2006.
2. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962-73.
3. Van der Laan M, Robins JM. Unified Methods for Censored Longitudinal Data and Causality. New York: Springer, 2003.
4. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004;58:265-71.
5. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some of the regressors are not always observed. Journal of the American Statistical Association 1994;89:846-66.
6. Scharfstein DO, Rotnitzky A, Robins J. Adjusting for non-ignorable drop-out using semiparametric non-response models - Comments & Rejoinder. Journal of the American Statistical Association 1999;94:1121-46.
7. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology 2009;20:3-5.