

Implementing double-robust estimators of causal effects

Richard Emsley

Biostatistics, Health Methodology Research Group
The University of Manchester, UK
richard.emsley@manchester.ac.uk

Mark Lunt

Arthritis Research Campaign Epidemiology Unit
The University of Manchester, UK

Andrew Pickles

Biostatistics, Health Methodology Research Group
The University of Manchester, UK

Graham Dunn

Biostatistics, Health Methodology Research Group
The University of Manchester, UK

Abstract. This article describes the implementation of a double-robust estimator for pretest–posttest studies (Lunceford and Davidian, 2004, *Statistics in Medicine* 23: 2937–2960) and presents a new Stata command (`dr`) that carries out the procedure. A double-robust estimator gives the analyst two opportunities for obtaining unbiased inference when adjusting for selection effects such as confounding by allowing for different forms of model misspecification; a double-robust estimator also can offer increased efficiency when all the models are correctly specified. We demonstrate the results with a Monte Carlo simulation study, and we show how to implement the double-robust estimator on a single simulated dataset, both manually and by using the `dr` command.

Keywords: st0149, dr, double-robust estimators, causal models, confounding, inverse probability of treatment weights, propensity score

1 Introduction

Pretest–posttest studies can be broadly defined and encompass a large number of situations ranging from prospective observational studies to randomized trials. Any study that measures outcome at the baseline and follows individuals over time before measuring outcome at the end of follow-up meets our definition of a pretest–posttest study. This would include longitudinal studies where the intermediate outcomes are not included in the analysis. We wish to carry out inference by comparing the final outcome between the two groups, while allowing for baseline values of the outcome and additional baseline covariates.

The ability to draw causal conclusions from observational data relies on the groups under consideration being exchangeable, such that they have identical distributions of characteristics; this is an immediate consequence from the counterfactual notion of causal effects. If exchangeability holds, it constrains associational measures (such as regression parameters) to be equal to the causal measures we wish to investigate.

In summary, the individual-level causal effect of treatment against control is defined by comparing subject A's outcome after receiving treatment with the same subject A's outcome had the subject received the control (the counterfactual outcome). As is intuitively obvious, these two outcomes cannot be observed simultaneously for subject A over the same time period. We therefore seek an alternative subject, subject B, with the same characteristics, for example, the same age and gender, and observe the outcome under the control exposure in this subject. We use this as a surrogate outcome for subject A's counterfactual outcome, allowing us to calculate an individual causal effect. We say that subject A and subject B are exchangeable.

Extending this to a group level, we wish to calculate the average causal effect. This gives the average effect for a randomly selected subject from the sample population. This is possible provided the two groups are exchangeable, such that the groups have identical distributions of variables. This is ensured in randomized controlled trials by randomization and in observational studies by thoughtful selection of the appropriate control group.

However, selection effects can invalidate the exchangeability assumption, and so the conclusions drawn from the associational measures are unlikely to be true causal relationships. Types of selection effect include missing data, nonadherence in a randomized study, and, the problem that is the focus in this article, confounding. Once the selection effects have been adjusted for, we make a stronger assumption of conditional exchangeability (conditional on these adjusted variables, exchangeability between the groups holds), which still allows us to equate the associational measures with the causal parameters we wish to estimate.

Confounding occurs when there are variables that are risk factors for the outcome of interest, correlated to the exposure under investigation, but which do not lie on the causal pathway between them. A variable meeting these criteria is called a confounder. A simple path diagram of confounding is shown in figure 1, showing the relationship between the exposure and the primary outcome being confounded by comorbidities.

(Continued on next page)

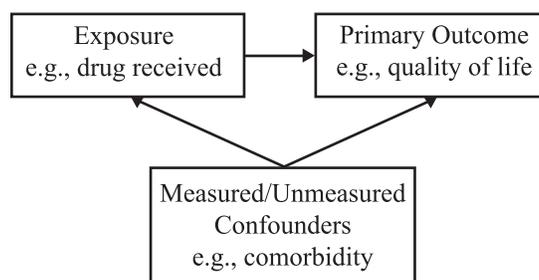


Figure 1. Confounding of the causal effect between exposure and the primary outcome

Confounding can be controlled for by using stratification, regression methods, or inverse probability weighting; we focus on the latter two. Regression analysis includes both the treatment variables and the confounders in the regression model as covariates, but we must specify the regression equation correctly for our resulting interpretation to be accurate. We say we have controlled for confounding while assuming *ceteris paribus*, with the coefficient of the treatment variable giving the average treatment effect.

The propensity score is a method that uses stratification, matching, or weighting to remove confounding. It is defined as the probability of receiving treatment given the measured covariates. For example, when treatment is binary, we can use a logistic or probit model with the baseline variables as covariates and take the predicted value from the regression as the subject's propensity score. Inverse probability weighting derives weights from the propensity score, where these are defined by the inverse of the propensity score if the subject receives treatment and the inverse of 1 minus the propensity score if the subject receives the control.

Any method that uses the propensity score requires that all relevant confounders are included in the model and that this model is specified correctly to validate the conditional exchangeability assumption. As with all observational studies, inference is only valid under the strong assumption of there being no unmeasured confounders. This is crucial to allowing the causal interpretation placed on the parameters but is, unfortunately, untestable.

Robins and colleagues (Bang and Robins 2005; Robins, Rotnitzky, and Zhao 1995; Robins 2000) introduced the concept of double-robust estimators that require a model for estimating the propensity score and the outcome model in the same estimator. The advantage these estimators offer is that they give unbiased estimates of the treatment effect when either one or both of these constituent models are correctly specified, thus allowing the analyst two opportunities for obtaining accurate results. The unmeasured confounders assumption is still required.

Lunceford and Davidian (2004) derived the form of double-robust estimators for pretest–posttest studies; this is what we implement in this article. The full details of the theory underpinning these estimators are accessible in the existing statistical literature

(Tsiatis 2006; Neugebauer and van der Laan 2005; Leon, Tsiatis, and Davidian 2003; van der Laan and Robins 2003). The theory is not covered in depth in this article, but there is a brief summary in section 2.

The remainder of this article is organized as follows: Section 2 introduces our notation, explains the concepts outlined in the introduction, and gives the definitions of the inverse probability of treatment-weighted (IPTW) estimators and the double-robust estimators. Section 3 explains how these estimators can be implemented manually in Stata before section 4 introduces the new `dr` command, which carries out the procedure, including the syntax and options. Section 5 presents the results of a simple simulation study demonstrating how the double-robust estimator offers increased protection against model misspecification, and section 6 offers a discussion of the work, including how these methods can be used in randomized trials.

2 Estimators

2.1 Notation

We introduce the notation we will use in the remainder of the article when describing the double-robust and IPTW estimators. We denote a subject, i , from a population, Ω , of size N to have received a binary exposure, A_i ($i = 1$ for treatment; $i = 0$ for control). Let $Y_{i,1}$ and $Y_{i,0}$ be the counterfactual posttest outcomes under treatment and control, respectively, and the individual causal effect is then defined as $Y_{i,1} - Y_{i,0}$. Which of these outcomes is actually observed depends entirely on the treatment variable A_i such that $Y_i = A_i Y_{i,1} + (1 - A_i) Y_{i,0}$, where Y_i denotes the observed continuous posttest outcome variable. We define the vector \underline{X}_i to include all baseline variables including the pretest outcome measurement.

The true average causal effect we wish to estimate in this context is $\tau = E(Y_{i,1} | \underline{X}_i) - E(Y_{i,0} | \underline{X}_i) = E(Y_i | \underline{X}_i, A_i = 1) - E(Y_i | \underline{X}_i, A_i = 0)$, with this last equality connecting the causal measures to the associational measure only valid under the exchangeability assumption.

2.2 IPTW estimators

All the estimates outlined in this section use the propensity score, which is defined as the probability of receiving treatment given the subjects' characteristics, i.e., $\pi_i = \pi_i(\underline{X}_i) = \Pr(A_i = 1 | \underline{X}_i)$. The propensity score was proposed by Rosenbaum and Rubin (1983) and can be used as a method for eliminating bias due to confounding. We use weighting by the estimated propensity score to generate inverse probability of treatment weights, where the estimated scores, \hat{p}_i , are predicted values from either a logistic or probit model.¹

1. Further information on computing and using propensity score methods in Stata can be found in Becker and Ichino (2002) and at <http://personalpages.manchester.ac.uk/staff/Mark.Lunt/> (click the link under Stata software and then go to the Propensity package).

Using simple inverse weights equal to $1/\hat{p}_i$ if $A_i = 1$ or $1/(1 - \hat{p}_i)$ if $A_i = 0$ leads to the following IPTW estimator (Lunceford and Davidian 2004):

$$\hat{\tau} = N^{-1} \sum_{i=1}^N \left(\frac{A_i Y_i}{\hat{p}_i} \right) - N^{-1} \sum_{i=1}^N \left\{ \frac{(1 - A_i) Y_i}{1 - \hat{p}_i} \right\}$$

This estimator averages over the total number of subjects in the population, but an alternative specification of IPTW estimator can be derived that averages over the sum of the weights for each group while still using the simple inverse weights, i.e.,

$$\hat{\tau}_{IPTW} = \left(\sum_{i=1}^N \frac{A_i}{\hat{p}_i} \right)^{-1} \sum_{i=1}^N \left(\frac{A_i Y_i}{\hat{p}_i} \right) - \left(\sum_{i=1}^N \frac{1 - A_i}{1 - \hat{p}_i} \right)^{-1} \sum_{i=1}^N \left\{ \frac{(1 - A_i) Y_i}{1 - \hat{p}_i} \right\}$$

This estimator, denoted by $\hat{\tau}_{IPTW}$, is that produced by the addition of the probability weight (`pweight`) in Stata and can be easily implemented, as we demonstrate in section 3. The following command would implement it within Stata: `regress Y A [pweight=wt]`, where `wt` is the inverse of the treatment probability defined using the propensity score.

To ensure the validity of the standard errors, a bootstrap procedure can be applied to the whole process including estimation of the propensity score. Alternatively, robust standard errors (of the Huber/White/sandwich context) are produced in Stata when weights are specified, and these also provide valid standard errors.

2.3 Double-robust estimators

We begin by briefly outlining the concept behind the double-robust estimators. The IPTW estimators belong to a class of unbiased estimating functions. By subtracting any term of expectation zero from the estimating equation, we still obtain unbiased estimates, and this additional term can be used to increase the efficiency of the estimators and provide protection against model misspecification. The estimator within this class with the smallest variance is called the semiparametric efficient estimator, and this is shown to be the double-robust estimator (Tsiatis 2006; Leon, Tsiatis, and Davidian 2003).

Lunceford and Davidian (2004) provide the following formula for a double-robust estimator in the pretest–posttest context:

$$\hat{\tau}_{DR} = \frac{1}{N} \sum_{i=1}^N \frac{A_i Y_i - (A_i - \hat{p}_i) m_1(\underline{X}_i)}{\hat{p}_i} - \frac{1}{N} \sum_{i=1}^N \frac{(1 - A_i) Y_i + (A_i - \hat{p}_i) m_0(\underline{X}_i)}{1 - \hat{p}_i}$$

where $m_A(\underline{X}_i) = E(Y_i | A_i = A, \underline{X}_i)$ for $A=0$ or $A=1$, i.e., these are the predicted values from regressions of the outcome on the baseline covariates, including the baseline outcome measure, where the coefficient estimates and predicted values are obtained from regressions carried out separately for each group but with the same model.

This estimator is known to be consistent for τ , the true average causal effect, if one of the following is true:

- The model for the propensity score is correctly specified, i.e., \hat{p} is the true propensity score.
- The regression model relating outcome to the baseline covariates is correctly specified, i.e., $m_A(\underline{X}_i)$ is equal to the true predicted values.

We can derive the robust form of the sampling variances; [Lunceford and Davidian \(2004\)](#) give this as $N^{-2} \sum_{i=1}^N \hat{I}_{\text{DR},i}^2$, where

$$\hat{I}_{\text{DR},i} = \frac{A_i Y_i - (A_i - \hat{p}_i) m_1(\underline{X}_i)}{\hat{p}_i} - \frac{(1 - A_i) Y_i + (A_i - \hat{p}_i) m_0(\underline{X}_i)}{1 - \hat{p}_i} - \tau_{\text{DR}}$$

Alternatively, we can bootstrap the procedure, including the estimation of the propensity scores, to obtain valid standard errors. This is an option in the `dr` command.

When both of the models are correctly specified, the double-robust estimator is the semiparametric efficient estimator. If the exposure model is correct, the double-robust estimator will have a smaller variance than the IPTW estimator. If the outcome model is correct, it will have a larger variance than the standard regression model, but it is offering protection against the misspecification of this model. When both the exposure and outcome models are misspecified, then the resulting estimate will be biased but, obviously, all alternative methods which also used these models, such as IPTW or regression, would also exhibit this bias.

3 Implementing double-robust estimators

We have previously introduced the method for calculating the IPTW estimator by using `pweights` in the regression command.

The double-robust estimators defined previously can be manually implemented by following the algorithm outlined below:

1. Fit a logistic (or probit) regression model for treatment conditional on the baseline variables (time-dependent variables can be included if required for longitudinal analysis). The predicted value from this regression gives the estimated propensity scores, \hat{p}_i .
2. Fit a regression model for outcome on the baseline variables for the treatment group only ($A_i = 1$), and obtain the predicted values for the whole sample. This gives the value for $\hat{m}_1(\underline{X}_i)$.
3. Fit the same regression model for outcome on the baseline variables for the control group only ($A_i = 0$), and obtain the predicted values for the whole sample. This gives the value for $\hat{m}_0(\underline{X}_i)$.

4. Substitute the predicted values \hat{p}_i , $\hat{m}_1(\underline{X}_i)$, and $\hat{m}_0(\underline{X}_i)$ into the expression for the double-robust estimator. This can be done by generating a new variable, which is then regressed as a constant value to ensure the averaging over N .

It is the specification of these models that gives the estimator its double-robust properties, but once specified, it is clear that the double-robust estimator in the pretest–posttest setting is simple to compute. The standard errors can be generated by bootstrapping the procedure, sampling at the level of the whole sequence of regressions, or by calculating the closed form of the formula in section 2. An example of this procedure is presented in section 5.

The obvious question when analyzing a real dataset is how the analyst knows whether his or her models are correctly specified; this is a concern that we emphasize in this article as essential to generating unbiased estimators. In reality, one cannot know and hence there is a vast literature on the statistical tests of model specification that can be applied. These tests and procedures can be used for building the outcome model, for example, fitting a stepwise `glm`, with the selected model being substituted into the `dr` command itself.

The model for the propensity score has a different purpose: we are interested in the predicted values from this model, not in making inference about the parameters. Advice on the propensity score recommends including as many variables as collected, even if they are not thought to be confounders, because it helps improve the prediction of the estimated propensity score. For example, [Lunceford and Davidian \(2004\)](#) derive theoretical results showing that including additional variables will increase the precision of the double-robust estimators, and they demonstrate this empirically through simulation. Extensions with additional variables to the simulation example in this article were carried out in [Emsley \(2007\)](#), and these also demonstrated efficiency gains under some misspecification combinations.

Including all possible variables in the exposure model should ensure that the estimated propensity scores are close to the true propensity score values for each subject, and it is this property which makes the double-robust estimator unbiased even under misspecification of the outcome model. When both models are correct, it also attains status as the most efficient estimator. However, the likely scenario is that both models are likely to be incorrectly specified, and so attempting to obtain the best possible estimated propensity scores is vital for the consistency of the estimator.

4 dr command

4.1 Syntax

```
dr outcome exposure [indepvars] [if] [in] [, ovars(varlist) pvars(varlist)
  family(familyname) link(linkname) vce(vce_type[, vce_options]) genvars
  suffix(string) ]
```

where *exposure* is binary and coded 0 for the control group and 1 for the treatment group. If necessary, use *xi* to generate *exposure*.

4.2 Options

ovars(*varlist*) determines which variables are used in the outcome-model part of the double-robust estimator. If it is not set, all the variables in *indepvars* are used.

pvars(*varlist*) determines which variables are used in the propensity-model part of the double-robust estimator. If it is not set, all the variables in *indepvars* are used.

family(*familyname*) uses the probability distribution of the generalized linear outcome model. See [R] **glm** for a list of alternatives.

link(*linkname*) is the link function of the generalized linear outcome model (see [R] **glm** for the options available). This only affects the outcome model: the double-robust estimator measures the difference between the mean outcome in the treated and the mean outcome in the untreated. So, if you have a dichotomous outcome, you may wish to specify **logit** for the link, but the double-robust estimate given will still be a risk difference, not an odds ratio.

vce(*vce_type* [, *vce_options*]) calculates the standard error of the estimator. *vce_type* can be either **robust** or **bootstrap**, with the default being *vce*(**robust**). The *vce*(**bootstrap**) option has a number of suboptions to control how the bootstrapping is performed; see [R] **bootstrap** for details. The options **oim** and **opg**, described in [R] *vce_option*, are not implemented. To reproduce a bootstrap analysis, set the *seed*(#) as outlined within *vce_options*. If this option is given, the random-number seed is set to the given value.

genvars adds the following variables to the dataset:

ptreat	Probability of receiving treatment, given <i>indepvars</i> or <i>pvars</i> () (if specified)
iptwt	Inverse probability of treatment weight
mu0	Predicted value of outcome in the untreated, given <i>indepvars</i> or <i>ovars</i> () (if specified)
mu1	Predicted value of outcome in the treated, given <i>indepvars</i> or <i>ovars</i> () (if specified)
mudiff	Difference between mu1 and mu0 in this observation

`suffix(string)` adds the suffix *string* to the variables produced by `genvars` to avoid potential name clashes.

4.3 Saved results

`dr` saves the following in `e()`:

Scalars	
<code>e(dr_est)</code>	double-robust estimate
<code>e(dr_var)</code>	variance of double-robust estimate
<code>e(dr0)</code>	expected average outcome if no subjects treated
<code>e(dr1)</code>	expected average outcome if all subjects treated
Macros	
<code>e(properties)</code>	"b V"
Matrices	
<code>e(b)</code>	coefficient vector
<code>e(V)</code>	variance-covariance matrix of the estimators
Functions	
<code>e(sample)</code>	marks estimation sample

When the `vce(bootstrap)` option is specified, additional results are returned; these are described under [R] **bootstrap**.

5 Simulation study

We performed a Monte Carlo simulation study to examine the ability of the double-robust estimator to deal with model misspecification and to test for potential efficiency gains over the IPTW approach.² We performed 5,500 simulations containing 1,000 subjects, and the results of these are shown in this section. The objectives of the study were

- to compare the bias of the double-robust estimators against traditional methods under various combinations of model misspecification, and
- to compare the efficiency of the estimators.

We also include Stata output indicating how to implement the double-robust estimators on one simulated dataset by using the same data-generating models. The causal diagram of the data structure is shown in figure 2.

2. This simulation study was part of a larger program of simulation work testing double-robust estimators in a variety of study designs and different sample sizes (Emsley 2007) based on simulation work by Davidian, Tsiatis, and Leon (2005). We present this study here to elucidate the ideas presented.

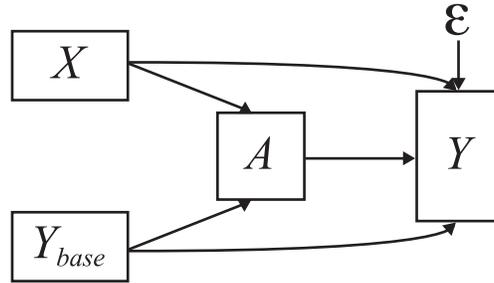


Figure 2. Causal diagram of the data in the simulation study

After generating the baseline binary variable X , which took the value 1 for the first half of the subjects and 0 for the remaining half, the baseline outcome variable Y_{base} was generated from a standard normal distribution. This gives $\underline{X} = (X, Y_{\text{base}})$, where $Y_{\text{base}} \sim N(0, 1)$ and $X = \begin{cases} 0 & \text{if } n > N/2 \\ 1 & \text{if } n \leq N/2 \end{cases}$.

The model for generating the true propensity score is

$$\pi = \Pr(A = 1 | \underline{X}) = \frac{1 - X}{1 + \exp(-1 + 0.1Y_{\text{base}})} + \frac{X}{1 + \exp(-0.9 - 0.1Y_{\text{base}})}$$

To generate a binary treatment, the probabilities predicted from the exposure model are compared to values drawn from a uniform (0,1) distribution. $A=0$ if π is less than the uniform variable, and $A=1$ otherwise.

The outcome variable Y is generated by linear and quadratic combinations of the baseline variables \underline{X} and the exposure A :

$$Y = 0.5A - 0.45 + 0.35Y_{\text{base}} + 0.1X + 0.3XY_{\text{base}} + 0.3XY_{\text{base}}^2 + 0.25Y_{\text{base}}^2 + \varepsilon$$

This model demonstrates that the true effect of the exposure β is the coefficient of A , in this instance equal to $\tau=0.5$.

We now explain the estimators we will use to compare the bias and efficiency of the double-robust estimators against the traditional methods. We begin by fitting the true model for Y , $E(Y | \underline{X}) = \alpha_0 + \alpha_1 A + \alpha_2 X + \alpha_3 Y_{\text{base}} + \alpha_4 XY_{\text{base}} + \alpha_5 Y_{\text{base}}^2 + \alpha_6 XY_{\text{base}}^2$, where α_1 is the causal effect estimator; we denote this as the true effect. A simple t test comparing Y between the two exposure groups, without accounting for any of the confounding baseline variables, gives a demonstration of the bias of the most naïve estimator.

To compute the remaining estimators, we need to specify the models for the propensity score and the outcome models for each group. These are shown below, and table 1 shows how we combine these to assess the effect of model misspecification.

True outcome: $Y = \alpha_0 + \alpha_1 A + \alpha_2 X + \alpha_3 Y_{\text{base}} + \alpha_4 XY_{\text{base}} + \alpha_5 Y_{\text{base}}^2 + \alpha_6 XY_{\text{base}}^2$

False outcome: $Y = \alpha_0^* + \alpha_1^* A + \alpha_2^* X + \alpha_3^* Y_{\text{base}} + \alpha_4^* XY_{\text{base}}$

True exposure: $\text{logit}(\pi) = \beta_0 + \beta_1 X + \beta_2 Y_{\text{base}} + \beta_3 XY_{\text{base}}$

False exposure: $\text{logit}(\pi) = \beta_0^* + \beta_1^* X + \beta_2^* Y_{\text{base}}$

Table 1. Models for generating predictors in the estimators

Estimator	Exposure model	Outcome model
True model	-	True
t test	-	-
IPTW	True	-
DR1	True	True
DR2	False	True
IPTW_fal	False	-
DR3	True	False
ANCOVA_fal	-	False
DR4	False	False
Wei ANCOVA_fal	False	False

We fit the correct exposure model, which includes the interaction term, and substitute the predicted values into the formula in section 2.2 to generate the IPTW estimator. We wish to compare this with the double-robust estimator with both exposure and outcome models, so, in the first double-robust estimator, the correctly specified propensity model and the true model for the regression are used to give the predicted values.

For the second double-robust estimator, we then misspecify the exposure model by ignoring the interaction of X and Y , and we take new predicted values as the propensity score. These values are also substituted into the IPTW formula for comparison, giving the IPTW_fal estimator.

In computing the third double-robust estimator, we use the correct exposure model and misspecify the outcome model by dropping the quadratic terms as covariates, and then we substitute the predicted values from this regression into the formula. This model is also fitted for the whole sample, and the value of α_1^* is taken as an estimate for comparison, denoted ANCOVA_fal. This then is just a standard regression adjustment for confounding but with a misspecified regression model.

Finally, we use the values from the misspecified exposure and outcome models together, substituting into the double-robust formula to give the final double-robust estimator. The theory tells us that this should not give unbiased estimators, and so computing it allows us to examine the degree of bias. We can compare it to a weighted regression (wei ANCOVA_fal), whereby we fit the misspecified outcome model using `pweights` with the incorrect propensity scores.

The standard errors of all the estimators are generated by bootstrapping the estimation procedures through 200 replications, with all the estimators defined together in one user-written program. This program is then bootstrapped, which allows for estimation of the propensity score each time and ensures that the estimators are compared on each bootstrap sample. Examples of how standard errors can be derived are shown on the individual simulation later in this section.

Table 2 gives the summary statistics for the 5,500 simulations, with the second column showing the Monte Carlo average for the estimators and the third column showing the bias relative to the true value of 0.5.³ We observe the average value of the true model to be 0.4991, and we see that the first double-robust estimator (DR1) gives the same value, compared with the IPTW effect estimate of 0.4976. As expected, because these estimators make use of the correctly specified exposure and outcome models, the estimates do not display significant bias. The bias of the *t* test, a naïve approach because it does not adjust for confounding, is clearly demonstrated.

Table 2. Summary statistics for simulation study

Estimator	MC mean	RB	MCS D	Boot SE	MSE ratio	CP (%)
True model	0.4991	-0.18	0.0702	0.0707	1.00	95.15
<i>t</i> test	0.3352	-32.96	0.0821	0.0817	0.15	47.55
IPTW	0.4976	-0.48	0.0820	0.0815	0.76	94.80
DR1	0.4991	-0.18	0.0705	0.0711	0.99	95.11
DR2	0.4991	-0.18	0.0705	0.0711	0.99	95.11
IPTW_fal	0.5120	+2.40	0.0834	0.0830	0.71	94.49
DR3	0.4973	-0.54	0.0836	0.0831	0.73	94.60
ANCOVA_fal	0.5135	+2.70	0.0836	0.0831	0.70	94.53
DR4	0.4965	-0.70	0.0833	0.0827	0.74	94.65
Wei ANCOVA_fal	0.5131	+2.62	0.0834	0.0830	0.71	94.40

Pairwise comparisons can be made between the double-robust estimators and their respective standard estimators to assess the effect of model misspecification. The second double-robust estimator and IPTW_fal have a misspecified exposure model when generating the propensity scores, but we observe that, while the double-robust estimator gives an unbiased estimate, the relative bias of IPTW_fal is shown in column 2. Similarly, the third double-robust estimator, which has a misspecified outcome model, performs better than ANCOVA_fal with the same model. The final double-robust estimator and weighted ANCOVA_fal have both exposure and outcome models misspecified, and the double-robust estimator displays less bias than the standard estimator.

3. In the table, MC mean is the Monte Carlo average; RB is the relative bias (%); MCS D is the Monte Carlo standard deviation; boot SE is the average of the bootstrapped standard errors (200 reps); MSE ratio is the mean square error for true divided by the mean square error of the indicated estimator; and CP is the empirical coverage probability of the 95% confidence interval.

We are able to produce two measures of the variability from the Monte Carlo simulations. The first is the Monte Carlo standard deviation, shown in column 3, which reflects the variability of the 5,500 effect estimates. Column 4 shows the average over the simulations of the bootstrapped standard errors, which is the error produced when analyzing one simulated dataset. We can see that the first two double-robust estimators are close to the most efficient estimator from the true model, as the theoretical results in section 2 predicted. Crucially, they are also more efficient than the respective IPTW estimators. When the outcome model is misspecified, the double-robust estimator does not offer gains in efficiency, but it is unbiased. Efficiency and bias together are assessed in the MSE ratio column.

Having shown that these estimators have the double-robust property and offer efficiency gains over the IPTW estimators, we now show how they can be implemented on a single dataset by using the manual procedure outlined in section 3 and the `dr` command presented in section 4.

First, we generate the data by using the models described previously and giving a summary of the dataset:

```
. summarize x ybase pi a y
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	1000	.5	.5002502	0	1
ybase	1000	-.0251784	1.012985	-3.038522	3.256222
pi	1000	.5104298	.2221019	.2332167	.7864826
a	1000	.509	.5001691	0	1
y	1000	.2704975	1.381972	-2.931211	9.049674

We can plot histograms of the true propensity score, π , over the two treatment groups, as shown in figure 3:

```
. histogram pi, by(a)
```

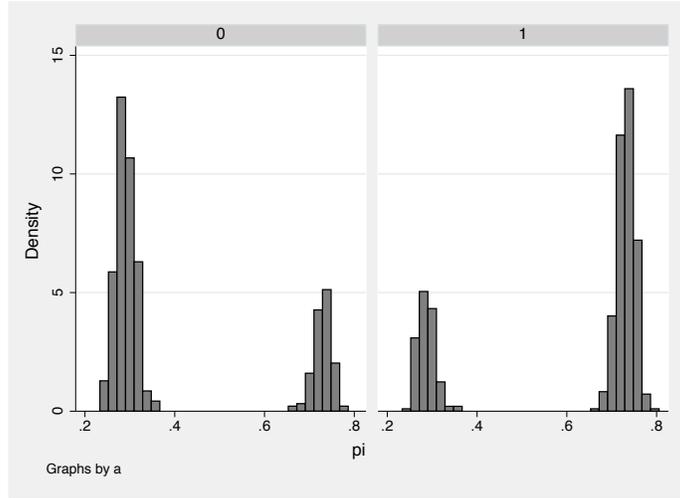


Figure 3. Histogram of true propensity scores by exposure group

The histograms show that the distributions of the propensity scores have similar ranges across the two groups, and the lack of spread across the middle of the range is because of the simple models used in this example. Importantly, the propensity scores do not have either extremely high or extremely low values, which can cause problems when inverse weighting by creating large, weighted outcome values (Kurth et al. 2006; Emsley 2007).

We now generate the estimates of the propensity score, which we label `pihat`, by taking the predicted values from a logistic regression of `A` on \underline{X} and then calculating the inverse weights. For this example, the correctly specified model for the propensity score is used.

```
. xi: logit a i.x*ybase
i.x          _Ix_0-1          (naturally coded; _Ix_0 omitted)
i.x*ybase    _IxXybase_#     (coded as above)
Iteration 0:  log likelihood = -692.98517
Iteration 1:  log likelihood = -579.10163
Iteration 2:  log likelihood = -578.37218
Iteration 3:  log likelihood = -578.37172
Logistic regression                                Number of obs =      1000
                                                    LR chi2(3)      =      229.23
                                                    Prob > chi2     =      0.0000
                                                    Pseudo R2      =      0.1654
Log likelihood = -578.37172
```

a	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ix_1	-2.023358	.1434155	-14.11	0.000	-2.304448	-1.742269
ybase	-.1404279	.1029371	-1.36	0.173	-.3421809	.0613251
_IxXybase_1	.021675	.1420704	0.15	0.879	-.2567778	.3001279
_cons	1.054866	.1024294	10.30	0.000	.8541081	1.255624

```
. predict pihat
(option p assumed; Pr(a))
. generate invwt=a/pihat + (1-a)/(1-pihat)
```

We can plot the values of `pihat` on a histogram to demonstrate the proximity to the real values of π ; see figure 4.

```
. histogram pihat, by(a)
```

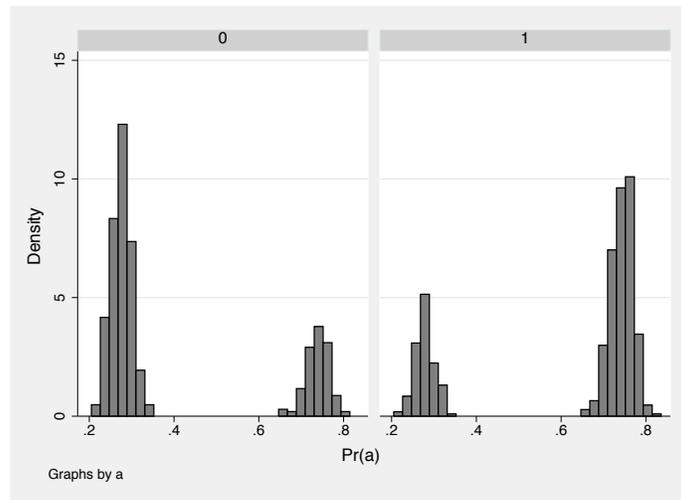


Figure 4. Histogram of estimated propensity scores by exposure group

The IPTW estimator can be generated by regressing Y on A with the inverse weights specified by using `pweights`.

```
. regress y a [pweight=invwt]
(sum of wgt is 1.9998e+03)
Linear regression Number of obs = 1000
F( 1, 998) = 26.95
Prob > F = 0.0000
R-squared = 0.0351
Root MSE = 1.3814
```

y	Robust		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
a	.5263717	.1013958	5.19	0.000	.3273983	.7253452
_cons	-.0196615	.0658265	-0.30	0.765	-.1488357	.1095127

This gives the average treatment effect of 0.5264 (SE=0.1014) in this dataset. We can compare this with the true value by fitting the correctly specified model, which gives the true effect of 0.5473 (SE=0.0716).

```
. regress y a x ybase intxy ysqr intxy2
(output omitted)
```

We then proceed to fit the correctly specified outcome models, from which we take the predicted values of $m_A(\underline{X})$ for $A=0$ and $A=1$.

```
. regress y ybase x intxy ysqr intxy2 if a==1
(output omitted)
. predict mu1
(option xb assumed; fitted values)
. regress y ybase x intxy ysqr intxy2 if a==0
(output omitted)
. predict mu0
(option xb assumed; fitted values)
```

Finally, we combine the predicted values of the propensity score and outcome values into the double-robust formula, and we summarize this value to generate the effect estimate. This is equivalent to the estimator DR1 from the simulation study, i.e., both the exposure and outcome models are correctly specified; we subsequently find the double-robust effect estimate is 0.5504, which is close to the value 0.5473 from the true model.

```
. generate mdiff1=-((a-pihat)*mu1/pihat)-((a-pihat)*mu0/(1-pihat))
. generate iptw=(2*a-1)*y*invtw
. generate dr1=iptw+mdiff1
. summarize dr1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
dr1	1000	.550427	2.174079	-11.15945	9.278258

The robust variance can be computed by subtracting the double-robust effect estimate from the subject's individual value, then squaring this and calculating the average, and finally dividing by the number of subjects.

```
. local dr_est=r(mean)
. tempvar I dr_var
. generate `I`=dr-`dr_est`
. generate I2=`I`^2
. summarize I2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
I2	1000	4.721894	10.78363	5.61e-06	137.1213

```
. generate `dr_var`=r(mean)/1000
. scalar dr_se=sqrt(`dr_var`)
. display dr_se
.06871604
```

The bootstrapped standard errors can be generated by bootstrapping the entire procedure, including the estimation of `pihat`. This is achieved by encoding the procedure within a written program and applying Stata's `bootstrap` command to this program.

Alternatively, the `dr` command we introduce in section 4 of this article will compute the robust standard errors as routine and the bootstrapped standard errors as an option. The command simply requires specification of the outcome variable, the exposure variable, and the exposure and outcome models. It also gives p -values and confidence intervals as standard. The command for generating the double-robust estimator with robust standard error is

```
. dr y a, ovars(ybase x intxy ysqr intxy2) pvars(x ybase intxy)
```

Doubly Robust Estimate of the effect of a on y
Using sandwich estimator of SE

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
a	.550427	.068716	8.01	0.000	.4157461	.685108

For the bootstrap with 200 replications, type

```
. dr y a, ovars(ybase x intxy ysqr intxy2) pvars(x ybase intxy)
> vce(bootstrap, reps(200) seed(1234))
```

Doubly Robust Estimate of the effect of a on y
(running dr on estimation sample)

Bootstrap replications (200)

	1	2	3	4	5
.....					50
.....					100
.....					150
.....					200

Bootstrap results

	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
a	.550427	.0720839	7.64	0.000	.4091452	.6917088

Number of obs = 1000
Replications = 200

As expected, the robust standard error is as we calculated manually, and the bootstrapped standard error is marginally larger. For both forms of standard error, the double-robust estimator is more efficient than the IPTW estimator. By adding the `genvars` option, the estimated propensity-score variable `ptreat` is added to the dataset. This can subsequently be used to generate histograms, as we produced earlier in this section, to check the distribution of the estimated propensity scores between the treatment and control groups and assess the region of common support.

We can also assess the degree of model misspecification in this simulated dataset by comparing the correlations between the predicted values from the correct and misspecified models. These show that the estimated propensity scores from the exposure model have a Pearson correlation of 1, indicating that they are identical. The estimate means of Y when $A=1$ under both forms of the outcome model have a correlation of 0.6689; when $A=0$ it is 0.7471 (both $p<0.001$). Although this applies only to the individual

dataset, replication of these values across the 5,500 simulations would account for the good performance of all the estimators shown in table 2.

6 Discussion

The most important assumption required when implementing the double-robust estimators, as with all observational studies, is that there are no unmeasured confounders. These are variables related to both exposure and outcome, and so if measured should be included in the models. Unfortunately, there is no way to empirically validate this assumption from the data. The emphasis should be on capturing as much information about potential confounders as possible from the study.

When using standard statistical methods, either for time-independent or time-varying treatments in a longitudinal setting such as marginal structural models, we would also require that our statistical models are correctly specified. The key property of the double-robust estimators is that we can weaken this assumption by giving the analyst two opportunities for correctly specifying either the exposure or outcome model. The reward for obtaining correct models for both is the increased efficiency that can be gained. The double-robust estimators can be used as a sensitivity analysis for assessing the standard statistical models used, because they should provide similar effect estimates if the standard models are correctly specified.

We propose the double-robust estimators in the context of a pretest–posttest study, applying the theoretical work of [Lunceford and Davidian \(2004\)](#). We emphasize that this can actually encompass a number of situations, including longitudinal studies where only the primary endpoint is being analyzed and the intermediate outcomes are not considered. For ease of exposition, we also framed the context of the article in an observational study; there are no statistical reasons for this. Randomized trials can be analyzed by using the same method, where the variables the analyst wishes to adjust for are those which, by chance, were not equally distributed between the two groups.

We have illustrated the ideas underpinning recent advances in causal analysis and highlighted the problem caused by selection effects. The double-robust estimators give two opportunities for adjusting for the hidden selection effects of confounding by combining inverse probability weighting with regression adjustment. Notably, the double-robust property is achieved by careful selection of the estimating equations; it is not a magical property ([Molenberghs 2005](#)).

The results of our simulation study clearly demonstrate the ability of the double-robust estimators to allow for model misspecification and the possible efficiency gains over IPTW estimators. While these have been established, we should highlight some limitations of our simulations. First, they reflect only one choice of parameter values and simple models. Second, because of the simplicity of the models, we used extreme forms of model misspecification by completely excluding variables known to be related to either exposure or outcome. Despite this, the predicted values from the exposure and outcome models, whether correctly or incorrectly specified, were highly correlated

with each other. Further simulation studies with alternative data-generating models would assess the extent of misspecification permissible. Application of the double-robust estimators to data from a clinical trial can be found in [Emsley \(2007\)](#).

We have shown how to implement the double-robust estimators manually through the simple use of standard Stata commands. The `dr` command we introduced in this article simplifies this procedure for the user, specifically with regard to obtaining valid standard errors for the estimates.

7 References

- Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962–973.
- Becker, S. O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2: 358–377.
- Davidian, M., A. A. Tsiatis, and S. Leon. 2005. Simulation study of estimators for treatment effect in a pretest-posttest study with posttest response missing at random. Working paper, Department of Statistics, North Carolina State University. <http://www4.stat.ncsu.edu/~davidian/sim.pdf>.
- Emsley, R. A. 2007. Statistical Models of Selection and Causation. PhD thesis, University of Manchester, UK.
- Kurth, T., A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, K. Berger, and J. M. Robins. 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 163: 262–270.
- Leon, S., A. A. Tsiatis, and M. Davidian. 2003. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* 59: 1046–1055.
- Lunceford, J. K., and M. Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23: 2937–2960.
- Molenberghs, G. 2005. Comment: Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* 20: 289–292.
- Neugebauer, R., and M. J. van der Laan. 2005. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* 129: 405–426.
- Robins, J. M. 2000. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10. Alexandria, VA: American Statistical Association.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90: 106–121.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Tsiatis, A. A. 2006. *Semiparametric Theory and Missing Data*. New York: Springer.

van der Laan, M. J., and J. M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

About the authors

Richard Emsley is a research associate in biomedical statistics in the Health Methodology Research Group at the University of Manchester. The research for this article was conducted during an MRC PhD Studentship, and he would like to acknowledge the support of the Medical Research Council in funding this work. His current research interests include statistical models of selection and causation, and the design and analysis of randomized trials of complex interventions in mental health.

Mark Lunt is a senior lecturer in biostatistics with the Arthritis Research Campaign Epidemiology Unit at the University of Manchester. His work largely involves the analysis of large observational studies, and he has a particular interest in causal analysis.

Andrew Pickles is a professor of epidemiological and social statistics at the University of Manchester. His research interests span multivariate statistics and include random effects, latent variable, causal, and developmental models of mixed data. He works primarily in the field of child mental health.

Graham Dunn is a professor of biomedical statistics and the head of the Health Methodology Research Group at the University of Manchester. He has particular interests in the evaluation of measurement errors and in causal inference in the presence of measurement error and selection effects. His main area of collaborative research is in clinical trials for the evaluation of cognitive behavior therapy and other psychological interventions.