Double robustness

Rhian M. Daniel

Division of Population Medicine, Cardiff University, 306c Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4YS.
Tel: +44 29 2068 7225. E-mail: DanielR8@cardiff.ac.uk.

Keywords: causal inference, convergence, data-adaptive estimation, inverse probability weighting, local efficiency, missing data, nuisance functionals, outcome regression.

Abstract

Double robustness is a prevalent topic in current statistical thinking, especially in causal inference and missing data methods. The most popular class of doubly robust (DR) estimators—also known as locally efficient augmented inverse probability weighted (AIPW) estimators—possess additional properties, linked to but distinct from double robustness, which lead to tractable statistical inferences and other attractive features.

In this article, we start by briefly introducing the concepts of nuisance models and multiple (double) robustness, before turning to a particular setting for a more formal introduction, namely that of estimating a population mean from incomplete data. For this setting, we describe the DR AIPW estimator and discuss the other properties alluded to above as well as double robustness itself. We then show how this setting is more general than it might seem at first sight, before going on to discuss many extensions, both to more complex settings and to recently-developed improved DR estimation strategies.
A Gentle Introduction: Nuisance Models and Multiple Robustness

Either by design or misfortune, statisticians must often model aspects of the distribution of their collected data that are not of primary scientific interest, as a step towards making inferences about the quantities of interest. For example, in an observational study, we must account for the potentially confounding effect of some covariates, not of intrinsic interest, in order to estimate the exposure–outcome effect of interest. Similarly, when analyzing incomplete data, it is judicious to account for variables that predict both the values of incomplete variables and the propensity to have them observed, even if these auxiliary variables would have been ignored had all the data been collected. Statistical models for the dependence of variables of interest on auxiliary variables, which are used as ingredients to estimate a parameter of interest, without being of intrinsic scientific interest, are called nuisance models.

Even when, as is standard in applied statistical practice in such settings, inference is based on a single regression model, a nuisance model is usually implicit. For example, suppose that in an observational study we have a binary exposure $X$, a continuous outcome $Y$ and a set of potential confounders $C = \{C_1, C_2, \ldots, C_p\}$, and we fit the single model

$$E(Y|X = x, C = c) = \gamma_{\text{int}} + \nu x + \gamma_1 c_1 + \gamma_2 c_2 + \cdots + \gamma_p c_p. \quad (1)$$

with $\nu$ the parameter of interest. In this setting (with binary exposure), the conditional expectation can (without any assumption) be written as

$$E(Y|X = x, C = c) = \eta_0(c) + \theta_0(c)x$$

where $\eta_0(c) = E(Y|X = 0, C = c)$ is a nuisance functional and $\theta_0(c) = E(Y|X = 1, C = c) - E(Y|X = 0, C = c)$ is the functional of interest. These can then be modeled using $\eta(c; \gamma)$ and $\theta(c; \nu)$, respectively, with the choice we illustrated in (1) being $\eta(c; \gamma) = \gamma_{\text{int}} + \gamma_1 c_1 + \gamma_2 c_2 + \cdots + \gamma_p c_p$ and $\theta(c; \nu) = \nu$, neither of which is necessarily correct.

Even when all other relevant assumptions hold, the misspecification of a nuisance model (such as $\eta(c; \gamma)$, above) typically (although not always) leads to bias and inconsistency in (and invalid inference about) the estimator of interest. Most commonly-used estimators, e.g. in our illustration above, make use of only one
nuisance model, and thus are reliant on specifying that model correctly. Some estimators make use of more than one nuisance model. A *multiply robust* estimator is one that remains consistent even if a subset of the nuisance models is incorrectly specified. In particular, a *doubly robust* (DR) estimator—the focus of this article—is one that makes use of two nuisance models, with the special property that the estimator is consistent as long as *at least one* of the two nuisance models is correctly specified. For example, our simple illustration above could be made DR by including the propensity score (the probability of exposure given confounders) on the right-hand side of (1): we return to this setting in Section 3.

Although the idea of having two ‘chances’ to specify a nuisance model correctly is intuitively appealing, the idea immediately raises some questions. For example, given that in most situations all models are wrong, do DR estimators outperform standard estimators when both nuisance models are incorrect? Also, DR estimators promise *consistency* when at least one model is correctly specified, but does this protection also extend to estimators of precision to deliver DR inference as well as DR estimation? Given the emphasis on model misspecification, can DR estimation be used in conjunction with data-adaptive/machine learning nuisance model estimation? We return to each of these issues in Section 3.

2 A more formal introduction using a simple example

2.1 Estimating a population mean from incomplete data under missing at random

Suppose we are interested in the population mean of a variable $Y$, and that we collect data from a randomly selected sample of size $n$ from this population. Suppose that $Y$ is observed only in a subset of the $n$ individuals, with $R = 1$ if $Y$ is observed, and $R = 0$ if $Y$ is missing. Suppose that covariates $X$ are observed on all $n$ individuals, and that the missing at random\(^{[1]}\) assumption is plausible given $X$, i.e. that $R \perp \!\!\!\perp Y \vert X$. Our task is therefore to make inference about the parameter of interest, $\mu_0 = E(Y)$, using the observed data $\{X_i, R_i, R_iY_i : i = 1, \ldots, n\}$ under the MAR assumption. We will assume throughout that $E(Y^2 \vert X = x) < \infty$ across the support of $X$.

Had all $n$ of the $Y$-values been observed, then we could ignore the covariate information and use:

$$\hat{\mu}^{\text{full}} = n^{-1} \sum_{i=1}^{n} Y_i.$$
However, unless the data are missing completely at random \((R \perp \perp Y)\), the corresponding complete records estimator,

\[
\hat{\mu}^{\text{CR}} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i},
\]

is biased and inconsistent as an estimator of \(\mu_0\), i.e. \(E(\hat{\mu}^{\text{CR}}) \neq \mu_0\) and \(\hat{\mu}^{\text{CR}} \not \xrightarrow{P} \mu_0\).

In this simple setting, missing data methods (see stat05051) make use of the fact that, under the MAR assumption, we can use \(X\) to ‘correct’ for the bias due to the incompleteness of the \(Y\) information\([2,3]\). This can be done in different ways, and involves nuisance models, as we now discuss. We start by describing estimators that rely on a single nuisance model, and from Section 2.4 onwards, we introduce doubly robust estimators.

Throughout, as well as MAR, we make the positivity assumption, namely that there exists an \(\varepsilon > 0\) such that, for all \(x\) in the support of \(X\), \(P(R = 1|X = x) > \varepsilon\). Colloquially, this says that for all ‘possible’ values of the covariates, the conditional probability of observing \(Y\) given these covariate values is bounded away from zero. Without this assumption, the ‘borrowing’ of \((X,Y)\)-information from the \(R = 1\) to the \(R = 0\) subgroups, which is the key idea underlying any MAR analysis, is not feasible, since not all the relevant information can be observed.

### 2.2 The Horvitz–Thompson inverse probability weighted estimator

The Horvitz–Thompson (see stat05701), or inverse probability weighted (IPW), estimator is\([3,4]\)

\[
\hat{\mu}^{\text{IPW}} = n^{-1} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi_0(X_i)}
\]  

where

\[
\pi_0(X) = P(R = 1|X).
\]

This is unbiased for \(\mu_0\) under MAR and positivity because:

\[
E \left\{ \frac{R Y}{\pi_0(X)} \right\} = E \left\{ \frac{Y}{\pi_0(X)} E (R|X,Y) \right\} = E \left\{ \frac{Y}{\pi_0(X)} E (R|X) \right\} = E \left\{ \frac{Y}{\pi_0(X)} \pi_0(X) \right\} = E(Y) = \mu_0
\]

where the antepenultimate equality follows from the MAR assumption. For consistency of \(\hat{\mu}^{\text{IPW}}\) we also require that the estimating function have finite variance, which is the case as long as positivity holds and
\( E(Y^2|x) < \infty \) across the support of \( X \), which we will assume throughout.

Since \( \pi_0(X) \) is typically an unknown function of \( X \), the estimator above is often called \textit{infeasible} (hence \( i \)-\( \text{IPW} \)). To turn it into a feasible estimator (\( f \)-\( \text{IPW} \)), \( \pi_0(X) \) must be estimated, typically by assuming that we know its functional form up to an unknown finite-dimensional parameter \( \alpha_0 \), so that \( \pi_0(X) = \pi(X; \alpha_0) \), where \( \pi(X; \alpha) \) is a known (i.e. assumed) function, smooth in \( \alpha \), and \( \alpha_0 \) is to be estimated from the data.

For example, we might specify a logistic regression model for \( R \) given \( X \) with main effects of each covariate only, with unknown coefficients \( \alpha_0 \):

\[
P(R = 1|X) = \pi(X; \alpha_0) = \frac{\exp\left(\alpha_0^T \tilde{X}\right)}{1 + \exp\left(\alpha_0^T \tilde{X}\right)}
\]

where \( \tilde{X} = (1, X^T)^T \). We could also choose more complex models (e.g. with higher-order terms), or indeed use a data-adaptive approach (see Section 3.4).

The model for \( R \) given \( X \) is an example of a nuisance model. After obtaining an estimate \( \hat{\alpha} \) of \( \alpha_0 \), e.g. by maximum likelihood, this can be used to form the feasible IPW estimator:

\[
\hat{\mu}_{f-\text{IPW}} = n^{-1} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})}.
\]

The class of densities \( M_\alpha \) is that permitted by our chosen model for \( R \) given \( X \), which in our example above is,

\[
M_\alpha = \{ \pi(X; \alpha) : \alpha \in \mathbb{R}^{p+1} \},
\]

where \( p \) is the dimension of \( X \). This model is correctly specified if and only if \( \pi_0(X) \in M_\alpha \), i.e. in our example if and only if there is some \( \alpha_0 \in \mathbb{R}^{p+1} \) such that \( \pi_0(X) = \pi(X; \alpha_0) \).

The estimator \( \hat{\mu}_{f-\text{IPW}} \) is consistent for \( \mu_0 \) under MAR and positivity, as long as our model \( M_\alpha \) is correctly specified and its parameters \( \alpha \) consistently estimated, as we now discuss.

\textbf{Aside: how the consistency of a feasible estimator follows from the consistency of the corresponding infeasible estimator.} \( \hat{\mu}_{f-\text{IPW}} \) is a consistent estimator of \( \mu_0 \) under the conditions listed above, because:

\[
\hat{\mu}_{f-\text{IPW}} = \hat{\mu}_{i-\text{IPW}} + n^{-1} \sum_{i=1}^{n} R_i Y_i \left\{ \frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi_0(X_i)} \right\}.
\]
\( \hat{\mu}^{\text{IPW}} \) converges in probability (see stat02847) to \( \mu_0 \) under MAR and positivity as we noted earlier, and thus the consistency of \( \hat{\mu}^{\text{IPW}} \) follows from the fact that

\[
n^{-1} \sum_{i=1}^{n} R_i Y_i \left\{ \frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi_0(X_i)} \right\}
\]

converges in probability to 0. This is so whenever \( M_\alpha \) is correctly specified and \( \alpha \) is consistently estimated, and essentially follows because of the continuity of \( g(x) = x^{-1} \) (as long as we exclude \( x = 0 \), which we can do under the positivity assumption) which allows us to argue that

\[
\pi(X, \hat{\alpha}) \xrightarrow{P} \pi_0(X) \Rightarrow \{\pi(X, \hat{\alpha})\}^{-1} \xrightarrow{P} \{\pi_0(X)\}^{-1}.
\]

For a more rigorous demonstration, see Tsiatis (2006).[3]

It is worth remembering the flavor of the argument above, since it is used repeatedly in the literature on this topic. The asymptotic behavior is usually first derived for the infeasible counterpart of the estimator in question, with the estimated nuisance functionals replaced by their probability limits. The asymptotic behavior of the feasible estimator is then inferred from this together with the behavior of the difference between the two estimators. The difference often converges in probability to zero, under certain assumptions, as was sketched above.

### 2.3 The outcome-regression estimator

In the previous subsection we gave an example of one possible estimator of \( \mu_0 \) that relies on a particular nuisance model, \( M_\alpha \), for the distribution of \( R \) given \( X \). We now discuss an alternative, based on specifying a nuisance model for the conditional mean of \( Y \) given \( X \) in the subset with \( R = 1 \).

The infeasible outcome regression estimator is

\[
\hat{\mu}^{\text{OR}} = n^{-1} \sum_{i=1}^{n} m_0(X_i)
\]

where

\[
m_0(X) = E(Y|X, R = 1).
\]

Note that \( m_0(x) \) is well-defined across the support of \( X \) under the positivity assumption. \( \hat{\mu}^{\text{OR}} \) is then unbiased and consistent for \( \mu_0 \) under MAR and positivity because

\[
E \{ m_0(X) \} = E \{ E(Y|X, R = 1) \} = E \{ E(Y|X) \} = E(Y) = \mu_0
\]

where the antepenultimate equality follows from MAR.
Since $m_0(X)$ is again typically an unknown function of $X$, it must be estimated, often by assuming that we know its functional form up to an unknown finite-dimensional parameter $\beta_0$, so that $m_0(X) = m(X; \beta_0)$, where $m(X; \beta)$ is a known (i.e. assumed) function, smooth in $\beta$, and $\beta_0$ is to be estimated from the data.

For example, we might specify a linear regression model for $Y$ given $X$ in the $R = 1$ subset, with main effects of each covariate only, with unknown coefficients $\beta_0$:

$$E(Y|X, R = 1) = m(X; \beta_0) = \beta^T \tilde{X}.$$  

The model for the mean of $Y$ given $X$ in the $R = 1$ subset is another example of a nuisance model, often called the outcome regression model. After obtaining an estimate $\hat{\beta}$ of $\beta_0$, e.g. by OLS, this can be used to form the feasible outcome-regression estimator:

$$\hat{\mu}_{f-OR} = n^{-1} \sum_{i=1}^{n} m(X_i; \hat{\beta}).$$

The class of densities $\mathcal{M}_\beta$ are those permitted by our chosen outcome regression model, in this example:

$$\mathcal{M}_\beta = \{ m(X; \beta) : \beta \in \mathbb{R}^{p+1} \}.$$  

This model is correctly specified if and only if $m_0(X) \in \mathcal{M}_\beta$, i.e. for this example if and only if there is some $\beta_0 \in \mathbb{R}^{p+1}$ such that $m_0(X) = m(X; \beta_0)$. Again, we could have chosen a different or more complex model, or indeed used a data-adaptive approach (see Section 3.4).

By a similar argument used in our aside above, the estimator $\hat{\mu}^{\text{f-OR}}$ is consistent for $\mu_0$ under MAR and positivity, as long as our model $\mathcal{M}_\beta$ is correctly specified and its parameters consistently estimated, e.g. by OLS. Among the class of (regular) estimators that are consistent under MAR, positivity and $\mathcal{M}_\beta$, $\hat{\mu}^{\text{f-OR}}$ asymptotically achieves the efficiency bound (see stat05838) when $\beta$ is estimated by OLS, by virtue of then being a maximum likelihood estimator.\(^{[5]}\) Typically, $\hat{\mu}^{\text{f-IPW}}$ is considerably less efficient than $\hat{\mu}^{\text{f-OR}}$.

### 2.4 Improving the efficiency of the IPW estimator

To improve the efficiency of $\hat{\mu}^{\text{f-IPW}}$, but without jeopardizing consistency under MAR, positivity and $\mathcal{M}_\alpha$, Robins et al\(^{[6]}\) considered a class of *augmented* IPW (AIPW) estimators, indexed by the augmentation
function $\phi(\cdot)$:

$$\hat{\mu}^{\phi-\text{AIPW}} = n^{-1} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi_0(X_i)} + \left\{ 1 - \frac{R_i}{\pi_0(X_i)} \right\} \phi(X_i) \right].$$

The choice $\phi(X) \equiv 0$ corresponds to the infeasible IPW estimator. For any choice of $\phi(\cdot)$, $\hat{\mu}^{\phi-\text{AIPW}}$ remains consistent for $\mu_0$ under MAR and positivity because:

$$E \left[ \left\{ 1 - \frac{R}{\pi_0(X)} \right\} \phi(X) \right] = E \left[ \phi(X) \left\{ 1 - \frac{E(R|X)}{\pi_0(X)} \right\} \right] = E \left[ \phi(X) \left\{ 1 - \frac{\pi_0(X)}{\pi_0(X)} \right\} \right] = 0 \quad \forall \phi(\cdot).$$

Some straightforward mathematics reveals that the choice of $\phi(\cdot)$ that yields the most efficient estimator in the above class is the $m_0$-AIPW estimator, given by $\phi(X) = m_0(X)$. Its feasible counterpart:

$$\hat{\mu}^{\ell-m-\text{AIPW}} = n^{-1} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left\{ 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right\} m(X_i; \hat{\beta}) \right]$$

makes use of two nuisance models, $\mathcal{M}_\alpha$ and $\mathcal{M}_\beta$.

This estimator ($\hat{\mu}^{\ell-m-\text{AIPW}}$) is consistent under MAR, positivity and $\mathcal{M}_\alpha$ even if $\mathcal{M}_\beta$ is incorrect. If $\mathcal{M}_\beta$ is correct, then it is the most efficient estimator in its class, namely the class of estimators consistent for $\mu_0$ under MAR, positivity and $\mathcal{M}_\alpha$. For a fuller, but accessible, account of these formal details, see Tsiatis (2006).[3]

2.5 Double robustness as a side-effect of local efficiency: two ways of viewing the same estimator

Turning back to the infeasible $m_0$-augmented estimator, note that it can be written in the following two ways:

$$\hat{\mu}^{\ell-m-\text{AIPW}} = n^{-1} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i)} + \left\{ 1 - \frac{R_i}{\pi(X_i)} \right\} m_0(X_i) \right] = n^{-1} \sum_{i=1}^{n} \left[ m_0(X_i) + \frac{R_i}{\pi_0(X_i)} \{ Y_i - m_0(X_i) \} \right].$$

The second way of writing it suggests that we also view it as a member of a $\psi$-augmented class of OR estimators:

$$\hat{\mu}^{\ell-\psi-\text{AOR}} = n^{-1} \sum_{i=1}^{n} \left[ m_0(X_i) + \frac{R_i}{\psi(X_i)} \{ Y_i - m_0(X_i) \} \right],$$
each member of which is consistent under MAR and positivity, since \( E \{ Y - m_0(X) \mid X \} = 0 \) and \( E \{ m_0(X) \} = \mu_0 \), irrespective of the choice of \( \psi(\cdot) \).

The choice \( \psi(X) = \pi_0(X) \) then corresponds to the one estimator in the class that, when we consider the feasible counterparts, remains consistent (under MAR and positivity) even when \( M_\beta \) is misspecified, as long as \( M_\alpha \) is correctly specified.

That is,

\[
\hat{\mu}^{f-m-\text{AIPW}} = n^{-1} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left( 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right) m(X_i; \hat{\beta}) \right] = \hat{\mu}^{f-\pi-\text{AOR}} = n^{-1} \sum_{i=1}^{n} \left[ m(X_i; \hat{\beta}) + \frac{R_i}{\pi(X_i; \hat{\alpha})} \{ Y_i - m(X_i; \hat{\beta}) \} \right]
\]

is consistent under MAR, positivity and \( M_\alpha \cup M_\beta \), i.e. if at least one of \( M_\alpha \) or \( M_\beta \) is correctly specified, but not necessarily both. This is the \textbf{double robustness} property, noted by Scharfstein et al.,\cite{7} although discussed earlier in the sample survey literature.\cite{8,9} From now on, we write f-SDR instead of either f-m-AIPW or f-\pi-AOR, where SDR stands for standard DR. ‘Standard’ is used to acknowledge that other DR estimators exist, other than the AIPW/AOR estimator introduced above, some of which will be discussed in Section 3.

We thus have two ways of viewing the same estimator: either (1) as the augmentation of IPW that preserves consistency (under \( M_\alpha \)) but improves efficiency (optimally under \( M_\alpha \cap M_\beta \)) and as a side-effect gains double robustness, i.e. consistency under just \( M_\alpha \cup M_\beta \), or (2) as the augmentation of OR that gains double robustness at the price of losing efficiency. In Figure 1, we include a heuristic visualization of the above.

### 2.6 Further intuition

For further intuition, see Figure 2. This Figure depicts the missing data setting when \( X \) is a continuous scalar random variable, so that \( Y \) against \( X \) can be plotted. Some individuals (those with \( R = 1 \)) have both \( X \) and \( Y \) observed, whereas others (those with \( R = 0 \)) have only \( X \) observed.

The red triangles on the left-hand side of Figure 2(a) represent the observed \( Y \)-values. A simple average of these would typically be biased as an estimator of \( \mu_0 \), but when weighted by the inverse probability weight \( 1/\pi(X) \), this weighted average will be consistent as long as the functional \( \pi(X) \) is correct, i.e. equal to \( \mathbb{P}(R = 1 \mid X) \).
On the other hand, we could proceed instead by outcome regression. That is, we would fit a model for the mean of \( Y \) given \( X \), represented by (in this example) the straight line through the observed \((X, Y)\) pairs. This allows us to evaluate \( m(X) \) both for the \( R = 1 \) and \( R = 0 \) individuals, and a simple average of the \( m(X) \), as shown on the right-hand side of Figure 2(a), is our OR estimator, consistent if our chosen linear regression model is correct.

We can now develop our intuition for the standard DR estimator, in particular by imagining it as an augmented OR estimator. This involves taking the \( m(X) \) predictions, leaving them untouched if \( R = 0 \), but ‘perturbing’ them by \( \pi(X)^{-1} \{ Y - m(X) \} \) if \( R = 1 \). If the linear regression model is already correctly specified, so that \( Y - m(X) \) has conditional mean zero given \( X \), then this introduces some additional residual variance but no bias.

However, if the \( m(X) \) functional is incorrect, as depicted in Figure 2(b), with the correct \( \mathbb{E}(Y|X) \) functional being the non-linear one as superimposed, then the DR estimator corrects for the bias in the \( m \)-functional, as long as the \( \pi \)-functional is correct. Intuitively, it does this by pulling each prediction that corresponds to an observed \( Y \)-value towards (and beyond) that observed \( Y \)-value. The extent to which it moves *beyond* the observed \( Y \)-value depends on the inverse weight; the bigger the weight, the bigger the shift. Also, the extent to which the prediction is perturbed depends on how far from the observed \( Y \)-value the prediction was, with the predictions corresponding to the largest residuals shifting the most. The fact that the points with the largest weights get the largest ‘over-correction’ intuitively makes sense since the \( R = 0 \) points remain uncorrected. The over-correction of the \( R = 1 \) predictions balances the under-correction of the \( R = 0 \) predictions, and the desired balance is achieved by reflecting the split between observed and unobserved \( Y \)-values at that level of \( X \), i.e. by using the (correct) \( \pi(X)^{-1} \).

### 2.7 Convergence and statistical inference

#### 2.7.1 Influence functions

The asymptotic behavior of (regular) estimators is governed by their *influence functions* (see stat01619). In order to discuss some of the other desirable properties (besides double robustness) of DR estimators, we briefly introduce some of this theory. We give only a rough sketch here, and recommend Tsiatis (2006)\[3\] for a much more detailed, and yet accessible, account.

Most reasonable estimators \( \hat{\mu} \) of a scalar parameter \( \mu_0 \) based on data \( \{ Z_i : i = 1, \ldots, n \} \) (so in our example,
$Z_i = (X_i, R_i, R_i Y_i)$ satisfy the following:

$$\sqrt{n} (\hat{\mu} - \mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi (Z_i) + o_P(1)$$  \hspace{1cm} (4)

where $o_P(1)$ denotes a remainder term that converges in probability to 0 (see stat02847), $E \{ \xi (Z) \} = 0$ and $0 < E \{ \xi (Z)^2 \} < \infty$. Then $\xi (Z)$ is known as the influence function of $\hat{\mu}$.

When (4) holds, it follows from the central limit theorem and Slutsky’s theorem that $\hat{\mu}$ is consistent and asymptotically normally distributed with:

$$\sqrt{n} (\hat{\mu} - \mu_0) \xrightarrow{D} N \left( 0, E \{ \xi (Z)^2 \} \right)$$  \hspace{1cm} (5)

where $\xrightarrow{D}$ denotes convergence in distribution.

Being able to write an estimator in the form (4) is useful, especially when the precise form of the $o_P(1)$ remainder term is complex, since it allows us to ignore this part when discussing the asymptotic behavior of the estimator, which is governed entirely by the influence function.

**2.7.2 The influence function of the feasible standard DR estimator**

Suppose that $\hat{\alpha} \xrightarrow{P} \alpha^*$ and $\hat{\beta} \xrightarrow{P} \beta^*$, where neither $\mathcal{M}_\alpha$ nor $\mathcal{M}_\beta$ is, for the time being, assumed correctly specified. Then, we can use (3) to write:

$$\sqrt{n} (\hat{\mu}^{\text{SDR}} - \mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left\{ 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right\} m(X_i; \hat{\beta}) - \mu_0 \right]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left\{ 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right\} m(X_i; \beta^*) - \mu_0 \right]$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ R_i \left\{ \frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi(X_i; \alpha^*)} \right\} \{Y_i - m(X_i; \beta^*)\} \right]$$

$$- \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \left\{ R_i - \pi(X_i; \alpha^*) \right\} \left\{ m(X_i; \hat{\beta}) - m(X_i; \beta^*) \right\} \right]$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ R_i \left\{ \frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi(X_i; \alpha^*)} \right\} \left\{ m(X_i; \hat{\beta}) - m(X_i; \beta^*) \right\} \right].$$

The first term (6) is $n^{-1/2}$ times the sum of the influence functions for a corresponding infeasible estimator.
(with the probability limits of $\hat{\alpha}$ and $\hat{\beta}$ used in place of these estimators).

The second term (7), under positivity, converges in probability to 0 as long as $\mathcal{M}_\beta$ is correctly specified, so that $m(\mathbf{X}; \beta^*) = m_0(\mathbf{X})$. Roughly speaking, this is because of the weak law of large numbers, which implies that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m_0(X_i)\} \xrightarrow{P} 0,$$

together with the fact that

$$\frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi(X_i; \alpha^*)}$$

is also converging (no matter how slowly). More formally, additional regularity conditions are needed in order to ensure uniform convergence of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m_0(X_i)\}$ to zero, so that (7) also converges to zero.\[3\]

Similarly, the third term (8), under positivity, converges in probability to 0 as long as $\mathcal{M}_\alpha$ is correctly specified, so that $\pi(\mathbf{X}; \alpha^*) = \pi_0(\mathbf{X})$, since then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{R_i - \pi_0(X_i)\} \xrightarrow{P} 0$$

by the weak law of large numbers.

Finally, if both $\hat{\alpha}$ and $\hat{\beta}$ are estimated by ML/OLS as described above, then both will converge at rate $O_P(n^{-1/2})$, meaning that the product

$$\left\{ \frac{1}{\pi(X_i; \hat{\alpha})} - \frac{1}{\pi(X_i; \alpha^*)} \right\} \left\{ m(\mathbf{X}_i; \hat{\beta}) - m(\mathbf{X}_i; \beta^*) \right\}$$

converges at rate $O_P(n^{-1})$. For the fourth term (9) to be $o_P(1)$, it suffices that (10) be $O_P(n^{-d})$ where $d > \frac{1}{2}$.

In summary, if both models are correctly specified ($\mathcal{M}_\alpha \cap \mathcal{M}_\beta$), and if the convergence rate for (10) is just faster than $\sqrt{n}$, e.g. if both $\hat{\alpha}$ and $\hat{\beta}$ converge at a rate just faster than $n^{1/4}$ each, then:

$$\sqrt{n} \left( \hat{\mu}^{fSDR} - \mu_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi_0(X_i)} + \left\{ 1 - \frac{R_i}{\pi_0(X_i)} \right\} m_0(X_i) - \mu_0 \right] + o_P(1)$$

and the influence function for the feasible and infeasible estimators coincide, which is a desirable property, as we describe in Sections 2.7.4 and 3.4.
2.7.3 Comparison with IPW and OR

Note that attempting a similar argument for the IPW or OR estimator would fail. In this case, the equivalent expression to (7)–(9) would consist of only one term, and for this to be \( o_P(1) \), we would need that \( \hat{\alpha} \) or \( \hat{\beta} \), respectively, converge at just faster than \( \sqrt{n} \) rate, i.e. faster than maximum likelihood.

Note that this does not mean that the IPW and OR estimators themselves are inconsistent, but rather that a naïve variance estimator (as described next) for them would be inconsistent.

2.7.4 Consequences for inference

The fact that the influence function for the standard DR estimator is the same as the influence function for its infeasible counterpart means that consistent estimation of the asymptotic variance of the standard DR estimator is straightforward. By (5), the asymptotic variance of \( \hat{\mu}^{f-\text{SDR}} \) is \( \frac{1}{n} \) times the variance of its influence function, and thus can be estimated using the sample variance of the influence functions, replacing each unknown parameter by its estimator, as follows:

\[
\hat{V}(\hat{\mu}^{f-\text{SDR}}) = \frac{1}{n^2} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left( 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right) m(X_i; \hat{\beta}) - \hat{\mu}^{f-\text{SDR}} \right]^2
\]

Under MAR, positivity, \( \mathcal{M}_\alpha \cap \mathcal{M}_\beta \) and sufficiently fast convergence of \( \hat{\alpha} \) and \( \hat{\beta} \) as outlined in Section 2.7.2, the variance estimator above is consistent. Furthermore, the asymptotic normality in (5) means that confidence intervals with asymptotically correct coverage can be similarly straightforwardly calculated.

We contrast this with other estimators, such as IPW and OR, which (as outlined in Section 2.7.3) lack this convenient property. Inference is thus more complicated for such estimators, whereby a contribution to the influence function resulting from the difference between the feasible and infeasible estimators must be incorporated. When the nuisance functionals correspond to parametric models, this is possible, although potentially cumbersome (see stat01678, and Stefanski and Boos\(^{[10]} \)). More generally, if data-adaptive estimation is used (see Section 3.4), then such calculations become analytically intractable, and even the bootstrap is invalid.\(^{[11]} \)

Note, however, that for a standard DR estimator, the easy ‘recipe’ outlined above (whereby an estimator’s variance is estimated using the sample variance of the estimated and easy-to-obtain influence functions for the corresponding infeasible estimator) is valid only when both nuisance models are correctly specified. Thus,
although this is a property inherently linked to the double robustness of the estimator, it is not itself a doubly robust property. See however one of our final remarks on this in Section 4.

3 Extensions and Further Issues

3.1 Standard DR estimators in other settings

The most immediate extension of the setting discussed in Section 2 is to the estimation of the average causal effect (ACE) in observational studies subject to confounding. Suppose $A$ denotes a binary exposure, $Y$ an outcome of interest, and $X$ a set of covariates. It is then typical to express the average causal effect of $A$ on $Y$ as $E\{Y(1) - Y(0)\}$ where $Y(a)$, $a = 0, 1$, is the potential value that $Y$ would take if $A$ were set by hypothetical intervention to the value $a$. $E\{Y(1)\}$ is then exactly analogous to $E(Y)$ in Section 2, with $R = A$ and $Y = Y(1)$; $\pi_0(X)$ is then the propensity score. By re-coding $R$, swapping 0 and 1, we also immediately obtain the standard DR estimator of $E\{Y(0)\}$, from which we obtain the standard DR of the ACE. Note that the combination of the positivity and MAR assumptions for both $R = 1$ and $R = 0$ leads to the usual positivity assumption in causal inference, together with the no unmeasured confounding assumption, that $A \perp \perp Y(a) | X$, for both $a = 0$ and $a = 1$. For more on this, see, for example, Bang and Robins.

Similar (standard DR) estimators, that are both augmented IPW and augmented OR estimators, have also been derived for longitudinal data (both for dealing with drop-out and the estimation of parameters of marginal structural models), natural direct and indirect effects in mediation analysis, for MAR incomplete covariates, for MNAR incomplete data, for time-to-event outcomes, for partially-observed composite outcomes, for attributable fractions, for the marginal probabilistic index (used in the Mann–Whitney test), in Randomized Controlled Trials, case–control studies, and in complex surveys.

3.2 DR estimators that are not AIPW/AOR

Not all DR estimators are augmented IPW and augmented OR estimators as described above. An important such class of alternative DR estimators are g-estimators (e.g. our very first example in Section 1 of regression adjustment for the propensity score) of structural mean and structural nested models. Examples include
estimators of statistical interaction parameters,\cite{25} controlled direct effects,\cite{26} average treatment effects in the treated in IV analyses,\cite{27–29} information-standardized effects,\cite{30} and conditional causal effects in longitudinal studies.\cite{31}

For other DR estimators in the setting of Section 2, see Kang and Schafer.\cite{32}

### 3.3 Improved DR estimators

Returning to the setting of Section 2 (for illustration), recall two properties of the standard DR estimator:

1. it is consistent (under MAR and positivity) when at least one, but not necessarily both, of $\mathcal{M}_\alpha$ and $\mathcal{M}_\beta$ is correctly specified;
2. it is asymptotically efficient, amongst estimators that are consistent under MAR, positivity and $\mathcal{M}_\alpha$, if $\mathcal{M}_\beta$ is correctly specified.

Upon reflecting on these two properties, it is natural to ask, respectively: (A) Is the bias of the standard DR estimator less than that of OR and IPW when both $\mathcal{M}_\alpha$ and $\mathcal{M}_\beta$ are misspecified? (B) Is standard DR always more efficient than IPW even when $\mathcal{M}_\beta$ is misspecified? In an influential paper in 2007, Kang and Schafer\cite{32} gave an example in which model misspecification is relatively mild, but the answers to (A) and (B) are both ‘no’.

This led to a number of proposed improvements to the standard DR estimator, in which the nuisance models are estimated in ways other than MLE/OLS. Many authors focused on (B) above, and developed nuisance model estimation strategies that guarantee improved efficiency relative to IPW even when $\mathcal{M}_\beta$ is incorrect.\cite{33–36}

In another proposal, focused on (A) above, Vermeulen and Vansteelandt (2015)\cite{37} develop bias-reduced DR estimation, which minimizes the asymptotic squared bias (in the direction of the nuisance parameters) when both models are locally misspecified. Their proposal has the nice property of preventing ‘bias inflation’, a phenomenon present in other DR estimators in which misspecification in either model is inflated at covariate values for which the estimated inverse probability weight is high.

van der Laan and colleagues have proposed another variant, namely targeted minimum loss-based estimation (TMLE). The key feature of TMLE is that it estimates the nuisance models in such a way that the resulting estimator is a substitution estimator, and thus cannot produce estimates outside the possible bounds of the parameter (e.g. risk differences outside [-1,1]). Note, however, that several of the other proposals\cite{35–37} share this property.
All of the above show substantial improvements on standard DR estimation in many numerical investigations. For a summary, see the chapter by Rotnitzky and Vansteelandt.\cite{Rotnitzky2005}

### 3.4 Data-adaptive estimation

Another consequence of the convergence result we sketched in Section 2.7.2, apart from leading to convenient inference, is that, unlike other estimators such as IPW and OR, good asymptotic properties of the resulting DR estimator can be achieved even when the convergence rates of the nuisance models is slower than the conventional parametric $\sqrt{n}$ rate. This opens the door to using data-adaptive (machine learning) estimation strategies to estimate $\pi_0(X)$ and $m_0(X)$ without incurring small sample bias, and retaining tractable inferences, as long as both estimated nuisance functionals converge to their respective truths, and that the convergence of the product shown in (10) is just faster than $\sqrt{n}$.\cite{Van der Laan2011}

This is why van der Laan and colleagues have based their data-adaptive TMLE on DR estimators.\cite{Van der Laan2013} Indeed, in recent promising work, Benkeser and van der Laan\cite{Benkeser2014} show that by including a particular data-adaptive algorithm (the highly adaptive lasso (HAL)), which converges at a rate just faster than $n^{1/4}$, in their SuperLearner ensemble machine learning approach,\cite{Benkeser2015} this property can in theory be guaranteed under a ‘bounded variation’ condition that is entirely plausible in most conceivable applications. In practice, however, it is likely that more restrictive conditions are needed for the HAL property to lead to good performance at realistic sample sizes.

Other authors are developing similar ideas\cite{Rosenbaum2015, van der Laan2016, van der Laan2017} in what is proving to be a very exciting area in high-dimensional causal inference at present.

### 4 Concluding Remarks

Double robustness is in itself an attractive property, since it weakens the conditions under which consistent estimation is achieved, and more generally offers a ‘compromise’ between competing estimators that rely on different nuisance model specifications. In particular, with the improvements mentioned in Section 3.3, DR estimators typically outperform competing estimators in a wide range of settings on a number of criteria.

In this article, we have reviewed further desirable properties of DR estimators, which may often prove more useful than double robustness itself, namely improved convergence (relative to other estimators) when
data-adaptive estimation strategies are employed for nuisance functional estimation, and tractable statistical inferences via the influence function that make DR estimators particularly appealing when variable selection and/or data-adaptive estimation techniques are employed.

Until recently, these additional properties required that both nuisance functionals converge to their respective truths. However, in recent work,\cite{43,45} this too has been relaxed. We can expect many similarly exciting developments in the near future.

5 Acknowledgements

The author is supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (grant 107617/Z/15/Z). The author is indebted to Stijn Vansteelandt for his comments on an earlier draft of this article, and more generally for his patient explanation of much of the technical aspects contained within it.

6 Related Articles

stat01619, stat01678, stat02847, stat05051, stat05701, stat05838.

7 References


8 Figures

Figure 1: A heuristic depiction of the standard DR estimator. We use it to picture how the efficiency of the estimator 

\[ n^{-1} \sum_{i=1}^{n} \left[ \frac{R_i Y_i}{\psi(X_i)} + \left\{ 1 - \frac{R_i}{\psi(X_i)} \right\} \phi(X_i) \right] \]

varies for different choices of \( \psi(\cdot) \) and \( \phi(\cdot) \). In addition, note that the estimator is in general only consistent along the two intersecting dotted lines that correspond to \( \psi(\cdot) = \pi_0(\cdot) \) OR \( \phi(\cdot) = m_0(\cdot) \). By considering the part of the surface above these two intersecting lines of consistency, efficiency is maximized at \( (\phi(\cdot) = m_0(\cdot), \psi(\cdot)^{-1} \equiv 0) \), i.e. the OR estimator. However, if we take only the slice through \( \psi(\cdot) = \pi_0(\cdot) \), then efficiency in that direction is maximized at \( (\phi(\cdot) = m_0(\cdot), \psi(\cdot) = \pi_0(\cdot)) \), i.e. the DR estimator. This illustrates the sense in which the standard DR estimator is locally, but not in general globally, efficient.
Figure 2: An example of the setting described in Section 2 with a single continuous covariate $X$. (a) shows the assumed linear model for $E(Y|X)$, whereas (b) superimposes the true non-linear $E(Y|X)$. This allows us to develop some intuition for the standard DR estimator both in the situations where the $Y|X$ model is correct and incorrect, as described in Section 2.6.